Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification
Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, Yi Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4692-4702
Recently, self-attention mechanisms have shown impressive performance in various NLP and CV tasks, which can help capture sequential characteristics and derive global information. In this work, we explore how to extend self-attention modules to better learn subtle feature embeddings for recognizing fine-grained objects, e.g., different bird species or person identities. To this end, we propose a dual cross-attention learning (DCAL) algorithm to coordinate with self-attention learning. First, we propose global-local cross-attention (GLCA) to enhance the interactions between global images and local high-response regions, which can help reinforce the spatial-wise discriminative clues for recognition. Second, we propose pair-wise cross-attention (PWCA) to establish the interactions between image pairs. PWCA can regularize the attention learning of an image by treating another image as distractor and will be removed during inference. We observe that DCAL can reduce misleading attentions and diffuse the attention response to discover more complementary parts for recognition. We conduct extensive evaluations on fine-grained visual categorization and object re-identification. Experiments demonstrate that DCAL performs on par with state-of-the-art methods and consistently improves multiple self-attention baselines, e.g., surpassing DeiT-Tiny and ViT-Base by 2.8% and 2.4% mAP on MSMT17, respectively.
****
SimAN: Exploring Self-Supervised Representation Learning of Scene Text via Similarity-Aware Normalization
Canjie Luo, Lianwen Jin, Jingdong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1039-1048
Recently self-supervised representation learning has drawn considerable attention from the scene text recognition community. Different from previous studies using contrastive learning, we tackle the issue from an alternative perspective, i.e., by formulating the representation learning scheme in a generative manner. Typically, the neighboring image patches among one text line tend to have similar styles, including the strokes, textures, colors, etc. Motivated by this common sense, we augment one image patch and use its neighboring patch as guidance to recover itself. Specifically, we propose a Similarity-Aware Normalization (SimAN) module to identify the different patterns and align the corresponding styles from the guiding patch. In this way, the network gains representation capability for distinguishing complex patterns such as messy strokes and cluttered backgrounds. Experiments show that the proposed SimAN significantly improves the representation quality and achieves promising performance. Moreover, we surprisingly find that our self-supervised generative network has impressive potential for data synthesis, text image editing, and font interpolation, which suggests that the proposed SimAN has a wide range of practical applications.
****
GASP, a Generalized Framework for Agglomerative Clustering of Signed Graphs and Its Application to Instance Segmentation
Alberto Bailoni, Constantin Pape, Nathan Hütsch, Steffen Wolf, Thorsten Beier, Anna Kreshuk, Fred A. Hamprecht; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11645-11655
We propose a theoretical framework that generalizes simple and fast algorithms for hierarchical agglomerative clustering to weighted graphs with both attractive and repulsive interactions between the nodes. This framework defines GASP, a Generalized Algorithm for Signed graph Partitioning, and allows us to explore many combinations of different linkage criteria and cannot-link constraints. We prove the equivalence of existing clustering methods to some of those combinations and introduce new algorithms for combinations that have not been studied before. We study both theoretical and empirical properties of these combinations and prove that some of these define an ultrametric on the graph. We conduct a systematic comparison of various instantiations of GASP on a large variety of both synthe

tic and existing signed clustering problems, in terms of accuracy but also effic iency and robustness to noise. Lastly, we show that some of the algorithms inclu ded in our framework, when combined with the predictions from a CNN model, resul t in a simple bottom-up instance segmentation pipeline. Going all the way from p ixels to final segments with a simple procedure, we achieve state-of-the-art acc uracy on the CREMI 2016 EM segmentation benchmark without requiring domain-speci fic superpixels.
****

Estimating Example Difficulty Using Variance of Gradients
Chirag Agarwal, Daniel D'souza, Sara Hooker; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10368-10378
In machine learning, a question of great interest is understanding what examples are challenging for a model to classify. Identifying atypical examples ensures the safe deployment of models, isolates samples that require further human inspe ction, and provides interpretability into model behavior. In this work, we propo se Variance of Gradients (VoG) as a valuable and efficient metric to rank data b y difficulty and to surface a tractable subset of the most challenging examples for human-in-the-loop auditing. We show that data points with high VoG scores ar e far more difficult for the model to learn and over-index on corrupted or memor ized examples. Further, restricting the evaluation to the test set instances wit h the lowest VoG improves the model's generalization performance. Finally, we sh ow that VoG is a valuable and efficient ranking for out-of-distribution detectio n
****

One Loss for Quantization: Deep Hashing With Discrete Wasserstein Distributional Matching
Khoa D. Doan, Peng Yang, Ping Li; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2022, pp. 9447-9457
Image hashing is a principled approximate nearest neighbor approach to find simi lar items to a query in a large collection of images. Hashing aims to learn a bi nary-output function that maps an image to a binary vector. For optimal retrieva l performance, producing balanced hash codes with low-quantization error to brid ge the gap between the learning stage's continuous relaxation and the inference stage's discrete quantization is important. However, in the existing deep superv ised hashing methods, coding balance and low-quantization error are difficult to achieve and involve several losses. We argue that this is because the existing quantization approaches in these methods are heuristically constructed and not e ffective to achieve these objectives. This paper considers an alternative approa ch to learning the quantization constraints. The task of learning balanced codes with low quantization error is re-formulated as matching the learned distributi on of the continuous codes to a pre-defined discrete, uniform distribution. This is equivalent to minimizing the distance between two distributions. We then pro pose a computationally efficient distributional distance by leveraging the discr ete property of the hash functions. This distributional distance is a valid dist ance and enjoys lower time and sample complexities. The proposed single-loss qua ntization objective can be integrated into any existing supervised hashing metho d to improve code balance and quantization error. Experiments confirm that the p roposed approach substantially improves the performance of several representativ e hashing methods.
****

Pixel Screening Based Intermediate Correction for Blind Deblurring
Meina Zhang, Yingying Fang, Guoxi Ni, Tieyong Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5892-590 0
Blind deblurring has attracted much interest with its wide applications in reali ty. The blind deblurring problem is usually solved by estimating the intermediat e kernel and the intermediate image alternatively, which will finally converge t o the blurring kernel of the observed image. Numerous works have been proposed t o obtain intermediate images with fewer undesirable artifacts by designing delic ate regularization on the latent solution. However, these methods still fail whi

le dealing with images containing saturations and large blurs. To address this p
roblem, we propose an intermediate image correction method which utilizes Bayes
posterior estimation to screen through the intermediate image and exclude those
unfavorable pixels to reduce their influence for kernel estimation. Extensive ex
periments have proved that the proposed method can effectively improve the accur
acy of the final derived kernel against the state-of-the-art methods on benchmar
k datasets by both quantitative and qualitative comparisons.
****

Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast
Ye Du, Zehua Fu, Qingjie Liu, Yunhong Wang; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4320-4329
Though image-level weakly supervised semantic segmentation (WSSS) has achieved g
reat progress with Class Activation Maps (CAMs) as the cornerstone, the large su
pervision gap between classification and segmentation still hampers the model to
 generate more complete and precise pseudo masks for segmentation. In this study
, we propose weakly-supervised pixel-to-prototype contrast that can provide pixe
l-level supervisory signals to narrow the gap. Guided by two intuitive priors, o
ur method is executed across different views and within per single view of an im
age, aiming to impose cross-view feature semantic consistency regularization and
 facilitate intra(inter)-class compactness(dispersion) of the feature space. Our
 method can be seamlessly incorporated into existing WSSS models without any cha
nges to the base networks and does not incur any extra inference burden. Extensi
ve experiments manifest that our method consistently improves two strong baselin
es by large margins, demonstrating the effectiveness. Specifically, built on top
 of SEAM, we improve the initial seed mIoU on PASCAL VOC 2012 from 55.4% to 61.5
%. Moreover, armed with our method, we increase the segmentation mIoU of EPS fro
m 70.8% to 73.6%, achieving new state-of-the-art.
****

Controllable Animation of Fluid Elements in Still Images
Aniruddha Mahapatra, Kuldeep Kulkarni; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3667-3676
We propose a method to interactively control the animation of fluid elements in
still images to generate cinemagraphs. Specifically, we focus on the animation o
f fluid elements like water, smoke, fire, which have the properties of repeating
 textures and continuous fluid motion. Taking inspiration from prior works, we r
epresent the motion of such fluid elements in the image in the form of a constan
t 2D optical flow map. To this end, we allow the user to provide any number of a
rrow directions and their associated speeds along with a mask of the regions the
 user wants to animate. The user-provided input arrow directions, their correspo
nding speed values, and the mask are then converted into a dense flow map repres
enting a constant optical flow map (F_D). We observe that F_D, obtained using si
mple exponential operations can closely approximate the plausible motion of elem
ents in the image. We further refine computed dense optical flow map F_D using a
 generative-adversarial network (GAN) to obtain a more realistic flow map. We de
vise a novel UNet based architecture to autoregressively generate future frames
using the refined optical flow map by forward-warping the input image features a
t different resolutions. We conduct extensive experiments on a publicly availabl
e dataset and show that our method is superior to the baselines in terms of qual
itative and quantitative metrics. In addition, we show the qualitative animation
s of the objects in directions that did not exist in the training set and provid
e a way to synthesize videos that otherwise would not exist in the real world. P
roject url: https://controllable-cinemagraphs.github.io/
****

Holocurtains: Programming Light Curtains via Binary Holography
Dorian Chan, Srinivasa G. Narasimhan, Matthew O'Toole; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17886
-17895
Light curtain systems are designed for detecting the presence of objects within
a user-defined 3D region of space, which has many applications across vision and
 robotics. However, the shape of light curtains have so far been limited to rule

d surfaces, i.e., surfaces composed of straight lines. In this work, we propose Holocurtains: a light-efficient approach to producing light curtains of arbitrary shape. The key idea is to synchronize a rolling-shutter camera with a 2D holographic projector, which steers (rather than block) light to generate bright structured light patterns. Our prototype projector uses a binary digital micromirror device (DMD) to generate the holographic interference patterns at high speeds. Our system produces 3D light curtains that cannot be achieved with traditional light curtain setups and thus enables all-new applications, including the ability to simultaneously capture multiple light curtains in a single frame, detect subtle changes in scene geometry, and transform any 3D surface into an optical touch interface.

****

Recurrent Dynamic Embedding for Video Object Segmentation

Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, Dong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1332-1341

Space-time memory (STM) based video object segmentation (VOS) networks usually keep increasing memory bank every several frames, which shows excellent performance. However, 1) the hardware cannot withstand the ever-increasing memory requirements as the video length increases. 2) Storing lots of information inevitably introduces lots of noise, which is not conducive to reading the most important information from the memory bank. In this paper, we propose a Recurrent Dynamic Embedding (RDE) to build a memory bank of constant size. Specifically, we explicitly generate and update RDE by the proposed Spatio-temporal Aggregation Module (SAM), which exploits the cue of historical information. To avoid error accumulation owing to the recurrent usage of SAM, we propose an unbiased guidance loss during the training stage, which makes SAM more robust in long videos. Moreover, the predicted masks in the memory bank are inaccurate due to the inaccurate network inference, which affects the segmentation of the query frame. To address this problem, we design a novel self-correction strategy so that the network can repair the embeddings of masks with different qualities in the memory bank. Extensive experiments show our method achieves the best tradeoff between performance and speed.

****

Deep Hierarchical Semantic Segmentation

Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1246-1257

Humans are able to recognize structured relations in observation, allowing us to decompose complex scenes into simpler parts and abstract the visual world in multiple levels. However, such hierarchical reasoning ability of human perception remains largely unexplored in current literature of semantic segmentation. Existing work is often aware of flatten labels and predicts target classes exclusively for each pixel. In this paper, we instead address hierarchical semantic segmentation (HSS), which aims at structured, pixel-wise description of visual observation in terms of a class hierarchy. We devise HSSN, a general HSS framework that tackles two critical issues in this task: i) how to efficiently adapt existing hierarchy-agnostic segmentation networks to the HSS setting, and ii) how to leverage the hierarchy information to regularize HSS network learning. To address i), HSSN directly casts HSS as a pixel-wise multi-label classification task, only bringing minimal architecture change to current segmentation models. To solve ii), HSSN first explores inherent properties of the hierarchy as a training objective, which enforces segmentation predictions to obey the hierarchy structure. Further, with hierarchy-induced margin constraints, HSSN reshapes the pixel embedding space, so as to generate well-structured pixel representations and improve segmentation eventually. We conduct experiments on four semantic segmentation datasets (i.e., Mapillary Vistas 2.0, Cityscapes, LIP, and PASCAL-Person-Part), with different class hierarchies, segmentation network architectures and backbones, showing the generalization and superiority of HSSN.

****

f-SfT: Shape-From-Template With a Physics-Based Deformation Model
Navami Kairanda, Edith Tretschk, Mohamed Elgharib, Christian Theobalt, Vladislav Golyanik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3948-3958
Shape-from-Template (SfT) methods estimate 3D surface deformations from a single monocular RGB camera while assuming a 3D state known in advance (a template). This is an important yet challenging problem due to the under-constrained nature of the monocular setting. Existing SfT techniques predominantly use geometric and simplified deformation models, which often limits their reconstruction abilities. In contrast to previous works, this paper proposes a new SfT approach explaining 2D observations through physical simulations accounting for forces and material properties. Our differentiable physics simulator regularises the surface evolution and optimises the material elastic properties such as bending coefficients, stretching stiffness and density. We use a differentiable renderer to minimise the dense reprojection error between the estimated 3D states and the input images and recover the deformation parameters using an adaptive gradient-based optimisation. For the evaluation, we record with an RGB-D camera challenging real surfaces exposed to physical forces with various material properties and textures. Our approach significantly reduces the 3D reconstruction error compared to multiple competing methods. For the source code and data, see https://4dqv.mpi-inf.mpg.de/phi-SfT/.
****
Continual Object Detection via Prototypical Task Correlation Guided Gating Mechanism
Binbin Yang, Xinchi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9255-9264
Continual learning is a challenging real-world problem for constructing a mature AI system when data are provided in a streaming fashion. Despite recent progress in continual classification, the researches of continual object detection are impeded by the diverse sizes and numbers of objects in each image. Different from previous works that tune the whole network for all tasks, in this work, we present a simple and flexible framework for continual object detection via pRotOtypical taSk corrElaTion guided gaTing mechAnism (ROSETTA). Concretely, a unified framework is shared by all tasks while task-aware gates are introduced to automatically select sub-models for specific tasks. In this way, various knowledge can be successively memorized by storing their corresponding sub-model weights in this system. To make ROSETTA automatically determine which experience is available and useful, a prototypical task correlation guided Gating Diversity Controller (GDC) is introduced to adaptively adjust the diversity of gates for the new task based on class-specific prototypes. GDC module computes class-to-class correlation matrix to depict the cross-task correlation, and hereby activates more exclusive gates for the new task if a significant domain gap is observed. Comprehensive experiments on COCO-VOC, KITTI-Kitchen, class-incremental detection on VOC and sequential learning of four tasks show that ROSETTA yields state-of-the-art performance on both task-based and class-based continual object detection.
****
DATA: Domain-Aware and Task-Aware Self-Supervised Learning
Qing Chang, Junran Peng, Lingxi Xie, Jiajun Sun, Haoran Yin, Qi Tian, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9841-9850
The paradigm of training models on massive data without label through self-supervised learning (SSL) and finetuning on many downstream tasks has become a trend recently. However, due to the high training costs and the unconsciousness of downstream usages, most self-supervised learning methods lack the capability to correspond to the diversities of downstream scenarios, as there are various data domains, latency constraints and etc. Neural architecture search (NAS) is one universally acknowledged fashion to conquer the issues above, but applying NAS on SSL seems impossible as there is no label or metric provided for judging model selection. In this paper, we present DATA, a simple yet effective NAS approach spec

ialized for SSL that provides Domain-Aware and Task-Aware pre-training. Specific ally, we (i) train a supernet which could be deemed as a set of millions of netw orks covering a wide range of model scales without any label, (ii) propose a fle xible searching mechanism compatible with SSL that enables finding networks of d ifferent computation costs, for various downstream vision tasks and data domains without explicit metric provided. Instantiated With MoCov2, our method achieves promising results across a wide range of computation costs on downstream tasks, including image classification, object detection and semantic segmentation. DAT A is orthogonal to most existing SSL methods and endows them the ability of cust omization on downstream needs. Extensive experiments on other SSL methods, inclu ding BYOL, ReSSL and DenseCL demonstrate the generalizability of the proposed me thod. Code would be made available soon.

****

TWIST: Two-Way Inter-Label Self-Training for Semi-Supervised 3D Instance Segment ation

Ruihang Chu, Xiaoqing Ye, Zhengzhe Liu, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, Jiay a Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 1100-1109

We explore the way to alleviate the label-hungry problem in a semi-supervised se tting for 3D instance segmentation. To leverage the unlabeled data to boost mode l performance, we present a novel Two-Way Inter-label Self-Training framework na med TWIST. It exploits inherent correlations between semantic understanding and instance information of a scene. Specifically, we consider two kinds of pseudo l abels for semantic- and instance-level supervision. Our key design is to provide object-level information for denoising pseudo labels and make use of their corr elation for two-way mutual enhancement, thereby iteratively promoting the pseudo -label qualities. TWIST attains leading performance on both ScanNet and S3DIS, c ompared to recent 3D pre-training approaches, and can cooperate with them to fur ther enhance performance, e.g., +4.4% AP50 on 1%-label ScanNet data-efficient be nchmark. Code is available at https://github.com/dvlab-research/TWIST.

****

Voxel Set Transformer: A Set-to-Set Approach to 3D Object Detection From Point C louds

Chenhang He, Ruihuang Li, Shuai Li, Lei Zhang; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8417-8427

Transformer has demonstrated promising performance in many 2D vision tasks. Howe ver, it is cumbersome to apply the self-attention underlying transformer on larg e-scale point cloud data because point cloud is a long sequence and unevenly dis tributed in 3D space. To solve this issue, existing methods usually compute self -attention locally by grouping the points into clusters of the same size, or per form convolutional self-attention on a discretized representation. However, the former results in stochastic point dropout, while the latter typically has narro w attention field. In this paper, we propose a novel voxel-based architecture, n amely Voxel Set Transformer (VoxSeT), to detect 3D objects from point clouds by means of set-to-set translation. VoxSeT is built upon a voxel-based set attentio n (VSA) module, which reduces the self-attention in each voxel by two cross-atte ntions and models features in a hidden space induced by a group of latent codes. With the VSA module, VoxSeT can manage voxelized point clusters with arbitrary size in a wide range, and process them in parallel with linear complexity. The p roposed VoxSeT integrates the high performance of transformer with the efficienc y of voxel-based model, which can be used as a good alternative to the convoluti onal and point-based backbones. VoxSeT reports competitive results on the KITTI and Waymo detection benchmarks. The source code of VoxSeT will be released.

****

Learning Adaptive Warping for Real-World Rolling Shutter Correction

Mingdeng Cao, Zhihang Zhong, Jiahao Wang, Yinqiang Zheng, Yujiu Yang; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17785-17793

This paper proposes a real-world rolling shutter (RS) correction dataset, BS-RSC , and a corresponding model to correct the RS frames in a distorted video. Mobil

e devices in the consumer market with CMOS-based sensors for video capture often result in rolling shutter effects when relative movements occur during the video acquisition process, calling for RS effect removal techniques. However, current state-of-the-art RS correction methods often fail to remove RS effects in real scenarios since the motions are various and hard to model. To address this issue, we propose a real-world RS correction dataset BS-RSC. Real distorted videos with corresponding ground truth are recorded simultaneously via a well-designed beam-splitter-based acquisition system. BS-RSC contains various motions of both camera and objects in dynamic scenes. Further, an RS correction model with adaptive warping is proposed. Our model can warp the learned RS features into global shutter counterparts adaptively with predicted multiple displacement fields. These warped features are aggregated and then reconstructed into high-quality global shutter frames in a coarse-to-fine strategy. Experimental results demonstrate the effectiveness of the proposed method, and our dataset can improve the model's ability to remove the RS effects in the real world.

****

Siamese Contrastive Embedding Network for Compositional Zero-Shot Learning

Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, Muli Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9326-9335

Compositional Zero-Shot Learning (CZSL) aims to recognize unseen compositions formed from seen state and object during training. Since the same state may be various in the visual appearance while entangled with different objects, CZSL is still a challenging task. Some methods recognize state and object with two trained classifiers, ignoring the impact of the interaction between object and state; the other methods try to learn the joint representation of the state-object compositions, leading to the domain gap between seen and unseen composition sets. In this paper, we propose a novel Siamese Contrastive Embedding Network (SCEN) for unseen composition recognition. Considering the entanglement between state and object, we embed the visual feature into a Siamese Contrastive Space to capture prototypes of them separately, alleviating the interaction between state and object. In addition, we design a State Transition Module (STM) to increase the diversity of training compositions, improving the robustness of the recognition model. Extensive experiments indicate that our method significantly outperforms the state-of-the-art approaches on three challenging benchmark datasets, including the recent proposed C-QGA dataset.

****

Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions

Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, Anima Anandkumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19056-19065

A significant gap remains between today's visual pattern recognition models and human-level visual cognition especially when it comes to few-shot learning and compositional reasoning of novel concepts. We introduce Bongard-HOI, a new visual reasoning benchmark that focuses on compositional learning of human-object interactions (HOIs) from natural images. It is inspired by two desirable characteristics from the classical Bongard problems (BPs): 1) few-shot concept learning, and 2) context-dependent reasoning. We carefully curate the few-shot instances with hard negatives, where positive and negative images only disagree on action labels, making mere recognition of object categories insufficient to complete our benchmarks. We also design multiple test sets to systematically study the generalization of visual learning models, where we vary the overlap of the HOI concepts between the training and test sets of few- shot instances, from partial to no overlaps. Bongard-HOI presents a substantial challenge to today's visual recognition models. The state-of-the-art HOI detection model achieves only 62% accuracy on few-shot binary prediction while even amateur human testers on MTurk have 91% accuracy. With the Bongard-HOI benchmark, we hope to further advance research efforts in visual reasoning, especially in holistic perception-reasoning systems and better representation learning.

****
RIM-Net: Recursive Implicit Fields for Unsupervised Learning of Hierarchical Shape Structures
Chengjie Niu, Manyi Li, Kai Xu, Hao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11779-11788
We introduce RIM-Net, a neural network which learns recursive implicit fields for unsupervised inference of hierarchical shape structures. Our network recursively decomposes an input 3D shape into two parts, resulting in a binary tree hierarchy. Each level of the tree corresponds to an assembly of shape parts, represented as implicit functions, to reconstruct the input shape. At each node of the tree, simultaneous feature decoding and shape decomposition are carried out by their respective feature and part decoders, with weight sharing across the same hierarchy level. As an implicit field decoder, the part decoder is designed to decompose a sub-shape, via a two-way branched reconstruction, where each branch predicts a set of parameters defining a Gaussian to serve as a local point distribution for shape reconstruction. With reconstruction losses accounted for at each hierarchy level and a decomposition loss at each node, our network training does not require any ground-truth segmentations, let alone hierarchies. Through extensive experiments and comparisons to state-of-the-art alternatives, we demonstrate the quality, consistency, and interpretability of hierarchical structural inference by RIM-Net.
****
Do Learned Representations Respect Causal Relationships?
Lan Wang, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 264-274
Data often has many semantic attributes that are causally associated with each other. But do attribute-specific learned representations of data also respect the same causal relations? We answer this question in three steps. First, we introduce NCINet, an approach for observational causal discovery from high-dimensional data. It is trained purely on synthetically generated representations and can be applied to real representations, and is specifically designed to mitigate the domain gap between the two. Second, we apply NCINet to identify the causal relations between image representations of different pairs of attributes with known and unknown causal relations between the labels. For this purpose, we consider image representations learned for predicting attributes on the 3D Shapes, CelebA, and the CASIA-WebFace datasets, which we annotate with multiple multi-class attributes. Third, we analyze the effect on the underlying causal relation between learned representations induced by various design choices in representation learning. Our experiments indicate that (1) NCINet significantly outperforms existing observational causal discovery approaches for estimating the causal relation between pairs of random samples, both in the presence and absence of an unobserved confounder, (2) under controlled scenarios, learned representations can indeed satisfy the underlying causal relations between their respective labels, and (3) the causal relations are positively correlated with the predictive capability of the representations. Code and annotations are available at: https://github.com/human-analysis/causal-relations-between-representations.
****
ZebraPose: Coarse To Fine Surface Encoding for 6DoF Object Pose Estimation
Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6738-6748
Establishing correspondences from image to 3D has been a key task of 6DoF object pose estimation for a long time. To predict pose more accurately, deeply learned dense maps replaced sparse templates. Dense methods also improved pose estimation in the presence of occlusion. More recently researchers have shown improvements by learning object fragments as segmentation. In this work, we present a discrete descriptor, which can represent the object surface densely. By incorporating a hierarchical binary grouping, we can encode the object surface very efficiently. Moreover, we propose a coarse to fine training strategy, which enables fine-grained correspondence prediction. Finally, by matching predicted codes with o

bject surface and using a PnP solver, we estimate the 6DoF pose. Results on the public LM-O and YCB-V datasets show major improvement over the state of the art w.r.t. ADD(-S) metric, even surpassing RGB-D based methods in some cases.
****

ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic
Yoad Tewel, Yoav Shalev, Idan Schwartz, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17918-17928
Recent text-to-image matching models apply contrastive learning to large corpora of uncurated pairs of images and sentences. While such models can provide a powerful score for matching and subsequent zero-shot tasks, they are not capable of generating caption given an image. In this work, we repurpose such models to generate a descriptive text given an image at inference time, without any further training or tuning step. This is done by combining the visual-semantic model with a large language model, benefiting from the knowledge in both web-scale models. The resulting captions are much less restrictive than those obtained by supervised captioning methods. Moreover, as a zero-shot learning method, it is extremely flexible and we demonstrate its ability to perform image arithmetic in which the inputs can be either images or text and the output is a sentence. This enables novel high-level vision capabilities such as comparing two images or solving visual analogy tests. Our code is available at: https://github.com/YoadTew/zero-shot-image-to-text.
****

Learning To Affiliate: Mutual Centralized Learning for Few-Shot Classification
Yang Liu, Weifeng Zhang, Chao Xiang, Tu Zheng, Deng Cai, Xiaofei He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14411-14420
Few-shot learning (FSL) aims to learn a classifier that can be easily adapted to accommodate new tasks, given only a few examples. To handle the limited-data in few-shot regimes, recent methods tend to collectively use a set of local features to densely represent an image instead of using a mixed global feature. They generally explore a unidirectional paradigm, e.g., find the nearest support feature for every query feature and aggregate these local matches for a joint classification. In this paper, we propose a novel Mutual Centralized Learning (MCL) to fully affiliate these two disjoint dense features sets in a bidirectional paradigm. We first associate each local feature with a particle that can bidirectionally random walk in a discrete feature space. To estimate the class probability, we propose the dense features' accessibility that measures the expected number of visits to the dense features of that class in a Markov process. We relate our method to learning a centrality on an affiliation network and demonstrate its capability to be plugged in existing methods by highlighting centralized local features. Experiments show that our method achieves the new state-of-the-art.
****

CAPRI-Net: Learning Compact CAD Shapes With Adaptive Primitive Assembly
Fenggen Yu, Zhiqin Chen, Manyi Li, Aditya Sanghi, Hooman Shayani, Ali Mahdavi-Amiri, Hao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11768-11778
We introduce CAPRI-Net, a self-supervised neural network for learning compact and interpretable implicit representations of 3D computer-aided design (CAD) models, in the form of adaptive primitive assemblies. Given an input 3D shape, our network reconstructs it by an assembly of quadric surface primitives via constructive solid geometry (CSG) operations. Without any ground-truth shape assemblies, our self-supervised network is trained with a reconstruction loss, leading to faithful 3D reconstructions with sharp edges and plausible CSG trees. While the parametric nature of CAD models does make them more predictable locally, at the shape level, there is much structural and topological variation, which presents a significant generalizability challenge to state-of-the-art neural models for 3D shapes. Our network addresses this challenge by adaptive training with respect to each test shape, with which we fine-tune the network that was pre-trained on a model collection. We evaluate our learning framework on both ShapeNet and ABC,

the largest and most diverse CAD dataset to date, in terms of reconstruction qua lity, sharp edges, compactness, and interpretability, to demonstrate superiority over current alternatives for neural CAD reconstruction.
****

ATPFL: Automatic Trajectory Prediction Model Design Under Federated Learning Fra mework
Chunnan Wang, Xiang Chen, Junzhe Wang, Hongzhi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6563-65 72
Although the Trajectory Prediction (TP) model has achieved great success in comp uter vision and robotics fields, its architecture and training scheme design rel y on heavy manual work and domain knowledge, which is not friendly to common use rs. Besides, the existing works ignore Federated Learning (FL) scenarios, failin g to make full use of distributed multi-source datasets with rich actual scenes to learn more a powerful TP model. In this paper, we make up for the above defec ts and propose ATPFL to help users federate multi-source trajectory datasets to automatically design and train a powerful TP model. In ATPFL, we build an effect ive TP search space by analyzing and summarizing the existing works. Then, based on the characters of this search space, we design a relation-sequence-aware sea rch strategy, realizing the automatic design of the TP model. Finally, we find a ppropriate federated training methods to respectively support the TP model searc h and final model training under the FL framework, ensuring both the search effi ciency and the final model performance. Extensive experimental results show that ATPFL can help users gain well-performed TP models, achieving better results th an the existing TP models trained on the single-source dataset.
****

Revisiting Learnable Affines for Batch Norm in Few-Shot Transfer Learning
Moslem Yazdanpanah, Aamer Abdul Rahman, Muawiz Chaudhary, Christian Desrosiers, Mohammad Havaei, Eugene Belilovsky, Samira Ebrahimi Kahou; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9 109-9118
Batch Normalization is a staple of computer vision models, including those emplo yed in few-shot learning. Batch Normalization layers in convolutional neural net works are composed of a normalization step, followed by a shift and scale of the se normalized features applied via the per-channel trainable affine parameters g amma and beta. These affine parameters were introduced to maintain the expressiv e powers of the model following normalization. While this hypothesis holds true for classification within the same domain, this work illustrates that these para meters are detrimental to downstream performance on common few-shot transfer tas ks. This effect is studied with multiple methods on well-known benchmarks such a s few-shot classification on miniImageNet and cross-domain few-shot learning (CD -FSL). Experiments reveal consistent performance improvements on CNNs with affin e unaccompanied Batch Normalization layers; particularly in large domain-shift f ew-shot transfer settings. As opposed to common practices in few-shot transfer l earning where the affine parameters are fixed during the adaptation phase, we sh ow fine-tuning them can lead to improved performance.
****

Bridging the Gap Between Classification and Localization for Weakly Supervised O bject Localization
Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, Sungroh Yoon; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 14258-14267
Weakly supervised object localization aims to find a target object region in a g iven image with only weak supervision, such as image-level labels. Most existing methods use a class activation map (CAM) to generate a localization map; howeve r, a CAM identifies only the most discriminative parts of a target object rather than the entire object region. In this work, we find the gap between classifica tion and localization in terms of the misalignment of the directions between an input feature and a class-specific weight. We demonstrate that the misalignment suppresses the activation of CAM in areas that are less discriminative but belon

g to the target object. To bridge the gap, we propose a method to align feature directions with a class-specific weight. The proposed method achieves a state-of-the-art localization performance on the CUB-200-2011 and ImageNet-1K benchmarks.

****

Multi-Class Token Transformer for Weakly Supervised Semantic Segmentation
Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Dan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4310-4319
This paper proposes a new transformer-based framework to learn class-specific object localization maps as pseudo labels for weakly supervised semantic segmentation (WSSS). Inspired by the fact that the attended regions of the one-class token in the standard vision transformer can be leveraged to form a class-agnostic localization map, we investigate if the transformer model can also effectively capture class-specific attention for more discriminative object localization by learning multiple class tokens within the transformer. To this end, we propose a Multi-class Token Transformer, termed as MCTformer, which uses multiple class tokens to learn interactions between the class tokens and the patch tokens. The proposed MCTformer can successfully produce class-discriminative object localization maps from the class-to-patch attentions corresponding to different class tokens. We also propose to use a patch-level pairwise affinity, which is extracted from the patch-to-patch transformer attention, to further refine the localization maps. Moreover, the proposed framework is shown to fully complement the Class Activation Mapping (CAM) method, leading to remarkably superior WSSS results on the PASCAL VOC and MS COCO datasets. These results underline the importance of the class token for WSSS.

****

3D Moments From Near-Duplicate Photos
Qianqian Wang, Zhengqi Li, David Salesin, Noah Snavely, Brian Curless, Janne Kontkanen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3906-3915
We introduce 3D Moments, a new computational photography effect. As input we take a pair of near-duplicate photos, i.e., photos of moving subjects from similar viewpoints, common in people's photo collections. As output, we produce a video that smoothly interpolates the scene motion from the first photo to the second, while also producing camera motion with parallax that gives a heightened sense of 3D. To achieve this effect, we represent the scene as a pair of feature-based layered depth images augmented with scene flow. This representation enables motion interpolation along with independent control of the camera viewpoint. Our system produces photorealistic space-time videos with motion parallax and scene dynamics, while plausibly recovering regions occluded in the original views. We conduct extensive experiments demonstrating superior performance over baselines on public datasets and in-the-wild photos. Project page: https://3d-moments.github.io/.

****

Exact Feature Distribution Matching for Arbitrary Style Transfer and Domain Generalization
Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8035-8045
Arbitrary style transfer (AST) and domain generalization (DG) are important yet challenging visual learning tasks, which can be cast as a feature distribution matching problem. With the assumption of Gaussian feature distribution, conventional feature distribution matching methods usually match the mean and standard deviation of features. However, the feature distributions of real-world data are usually much more complicated than Gaussian, which cannot be accurately matched by using only the first-order and second-order statistics, while it is computationally prohibitive to use high-order statistics for distribution matching. In this work, we, for the first time to our best knowledge, propose to perform Exact Feature Distribution Matching (EFDM) by exactly matching the empirical Cumulative

Distribution Functions (eCDFs) of image features, which could be implemented by applying the Exact Histogram Matching (EHM) in the image feature space. Particularly, a fast EHM algorithm, named Sort-Matching, is employed to perform EFDM in a plug-and-play manner with minimal cost. The effectiveness of our proposed EFDM method is verified on a variety of AST and DG tasks, demonstrating new state-of-the-art results. Codes are available at https://github.com/YBZh/EFDM.

****

Blind2Unblind: Self-Supervised Image Denoising With Visible Blind Spots
Zejin Wang, Jiazheng Liu, Guoqing Li, Hua Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2027-2036
Real noisy-clean pairs on a large scale are costly and difficult to obtain. Meanwhile, supervised denoisers trained on synthetic data perform poorly in practice. Self-supervised denoisers, which learn only from single noisy images, solve the data collection problem. However, self-supervised denoising methods, especially blindspot-driven ones, suffer sizable information loss during input or network design. The absence of valuable information dramatically reduces the upper bound of denoising performance. In this paper, we propose a simple yet efficient approach called Blind2Unblind to overcome the information loss in blindspot-driven denoising methods. First, we introduce a global-aware mask mapper that enables global perception and accelerates training. The mask mapper samples all pixels at blind spots on denoised volumes and maps them to the same channel, allowing the loss function to optimize all blind spots at once. Second, we propose a re-visible loss to train the denoising network and make blind spots visible. The denoiser can learn directly from raw noise images without losing information or being trapped in identity mapping. We also theoretically analyze the convergence of the re-visible loss. Extensive experiments on synthetic and real-world datasets demonstrate the superior performance of our approach compared to previous work. Code is available at https://github.com/demonsjin/Blind2Unblind.

****

Balanced and Hierarchical Relation Learning for One-Shot Object Detection
Hanqing Yang, Sijia Cai, Hualian Sheng, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Yong Tang, Yu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7591-7600
Instance-level feature matching is significantly important to the success of modern one-shot object detectors. Recently, the methods based on the metric-learning paradigm have achieved an impressive process. Most of these works only measure the relations between query and target objects on a single level, resulting in suboptimal performance overall. In this paper, we introduce the balanced and hierarchical learning for our detector. The contributions are two-fold: firstly, a novel Instance-level Hierarchical Relation (IHR) module is proposed to encode the contrastive-level, salient-level, and attention-level relations simultaneously to enhance the query-relevant similarity representation. Secondly, we notice that the batch training of the IHR module is substantially hindered by the positive-negative sample imbalance in the one-shot scenario. We then introduce a simple but effective Ratio-Preserving Loss (RPL) to protect the learning of rare positive samples and suppress the effects of negative samples. Our loss can adjust the weight for each sample adaptively, ensuring the desired positive-negative ratio consistency and boosting query-related IHR learning. Extensive experiments show that our method outperforms the state-of-the-art method by 1.6% and 1.3% on PASCAL VOC and MS COCO datasets for unseen classes, respectively. The code will be available at https://github.com/hero-y/BHRL.

****

End-to-End Generative Pretraining for Multimodal Video Captioning
Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17959-17968
Recent video and language pretraining frameworks lack the ability to generate sentences. We present Multimodal Video Generative Pretraining (MV-GPT), a new pretraining framework for learning from unlabelled videos which can be effectively used for generative tasks such as multimodal video captioning. Unlike recent vide

o-language pretraining frameworks, our framework trains both a multimodal video encoder and a sentence decoder jointly. To overcome the lack of captions in unlabelled videos, we leverage the future utterance as an additional text source and propose a bidirectional generation objective -- we generate future utterances given the present mulitmodal context, and also the present utterance given future observations. With this objective, we train an encoder-decoder model end-to-end to generate a caption from raw pixels and transcribed speech directly. Our model achieves state-of-the-art performance for multimodal video captioning on four standard benchmarks, as well as for other video understanding tasks such as generative and discriminative VideoQA, video retrieval and action classification.
****

Delving Deep Into the Generalization of Vision Transformers Under Distribution Shifts

Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7277-7286

Recently, Vision Transformers have achieved impressive results on various Vision tasks. Yet, their generalization ability under different distribution shifts is poorly understood. In this work, we provide a comprehensive study on the out-of-distribution generalization of Vision Transformers. To support a systematic investigation, we first present a taxonomy of distribution shifts by categorizing them into five conceptual levels: corruption shift, background shift, texture shift, destruction shift, and style shift. Then we perform extensive evaluations of Vision Transformer variants under different levels of distribution shifts and compare their generalization ability with Convolutional Neural Network (CNN) models. Several important observations are obtained: 1) Vision Transformers generalize better than CNNs under multiple distribution shifts. With the same or less amount of parameters, Vision Transformers are ahead of corresponding CNNs by more than 5% in top-1 accuracy under most types of distribution shift. In particular, Vision Transformers lead by more than 10% under the corruption shifts. 2) larger Vision Transformers gradually narrow the in-distribution (ID) and out-of-distribution (OOD) performance gap. To further improve the generalization of Vision Transformers, we design the enhanced Vision Transformers through self-supervised learning, information theory, and adversarial learning. By investigating these three types of generalization-enhanced Transformers, we observe the gradient-sensitivity of Vision Transformers and design a smoother learning strategy to achieve a stable training process. With modified training schemes, we achieve improvements on performance towards out-of-distribution data by 4% from vanilla Vision Transformers. We comprehensively compare these three types of generalization-enhanced Vision Transformers with their corresponding CNN models and observe that: 1) For the enhanced model, larger Vision Transformers still benefit more from the out-of-distribution generalization. 2) generalization-enhanced Vision Transformers are more sensitive to the hyper-parameters than their corresponding CNN models. We hope our comprehensive study could shed light on the design of more generalizable learning systems.
****

NICE-SLAM: Neural Implicit Scalable Encoding for SLAM

Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12786-12796

Neural implicit representations have recently shown encouraging results in various domains, including promising progress in simultaneous localization and mapping (SLAM). Nevertheless, existing methods produce over-smoothed scene reconstructions and have difficulty scaling up to large scenes. These limitations are mainly due to their simple fully-connected network architecture that does not incorporate local information in the observations. In this paper, we present NICE-SLAM, a dense SLAM system that incorporates multi-level local information by introducing a hierarchical scene representation. Optimizing this representation with pre-trained geometric priors enables detailed reconstruction on large indoor scenes. Compared to recent neural implicit SLAM systems, our approach is more scalable

, efficient, and robust. Experiments on five challenging datasets demonstrate co
mpetitive results of NICE-SLAM in both mapping and tracking quality. Project pag
e: https://pengsongyou.github.io/nice-slam
****

HyperDet3D: Learning a Scene-Conditioned 3D Object Detector
Yu Zheng, Yueqi Duan, Jiwen Lu, Jie Zhou, Qi Tian; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5585-5594
A bathtub in a library, a sink in an office, a bed in a laundry room - the count
er-intuition suggests that scene provides important prior knowledge for 3D objec
t detection, which instructs to eliminate the ambiguous detection of similar obj
ects. In this paper, we propose HyperDet3D to explore scene-conditioned prior kn
owledge for 3D object detection. Existing methods strive for better representati
on of local elements and their relations without sceneconditioned knowledge, whi
ch may cause ambiguity merely based on the understanding of individual points an
d object candidates. Instead, HyperDet3D simultaneously learns scene-agnostic em
beddings and scene-specific knowledge through scene-conditioned hypernetworks. M
ore specifically, our HyperDet3D not only explores the sharable abstracts from v
arious 3D scenes, but also adapts the detector to the given scene at test time.
We propose a discriminative Multi-head Scene-specific Attention (MSA) module to
dynamically control the layer parameters of the detector conditioned on the fusi
on of scene-conditioned knowledge. Our HyperDet3D achieves state-of-the-art resu
lts on the 3D object detection benchmark of the ScanNet and SUN RGB-D datasets.
Moreover, through cross-dataset evaluation, we show the acquired scene-condition
ed prior knowledge still takes effect when facing 3D scenes with domain gap.
****

Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion
Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, Jiwen
Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2022, pp. 17113-17122
Human behavior has the nature of indeterminacy, which requires the pedestrian tr
ajectory prediction system to model the multi-modality of future motion states.
Unlike existing stochastic trajectory prediction methods which usually use a lat
ent variable to represent multi-modality, we explicitly simulate the process of
human motion variation from indeterminate to determinate. In this paper, we pres
ent a new framework to formulate the trajectory prediction task as a reverse pro
cess of motion indeterminacy diffusion (MID), in which we progressively discard
indeterminacy from all the walkable areas until reaching the desired trajectory.
 This process is learned with a parameterized Markov chain conditioned by the ob
served trajectories. We can adjust the length of the chain to control the degree
 of indeterminacy and balance the diversity and determinacy of the predictions.
Specifically, we encode the history behavior information and the social interact
ions as a state embedding and devise a Transformer-based diffusion model to capt
ure the temporal dependencies of trajectories. Extensive experiments on the huma
n trajectory prediction benchmarks including the Stanford Drone and ETH/UCY data
sets demonstrate the superiority of our method. Code is available at https://git
hub.com/gutianpei/MID.
****

CLRNet: Cross Layer Refinement Network for Lane Detection
Tu Zheng, Yifei Huang, Yang Liu, Wenjian Tang, Zheng Yang, Deng Cai, Xiaofei He;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2022, pp. 898-907
Lane is critical in the vision navigation system of the intelligent vehicle. Nat
urally, lane is a traffic sign with high-level semantics, whereas it owns the sp
ecific local pattern which needs detailed low-level features to localize accurat
ely. Using different feature levels is of great importance for accurate lane det
ection, but it is still under-explored. In this work, we present Cross Layer Ref
inement Network (CLRNet) aiming at fully utilizing both high-level and low-level
 features in lane detection. In particular, it first detects lanes with high-lev
el semantic features then performs refinement based on low-level features. In th
is way, we can exploit more contextual information to detect lanes while leverag

ing local detailed lane features to improve localization accuracy. We present ROIGather to gather global context, which further enhances the feature representation of lanes. In addition to our novel network design, we introduce Line IoU loss which regresses the lane line as a whole unit to improve the localization accuracy. Experiments demonstrate that the proposed method greatly outperforms the state-of-the-art lane detection approaches. Code is available at:https://github.com/Turoad/CLRNet.
****

Cross-Modal Map Learning for Vision and Language Navigation
Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, Kostas Daniilidis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15460-15470
We consider the problem of Vision-and-Language Navigation (VLN). The majority of current methods for VLN are trained end-to-end using either unstructured memory such as LSTM, or using cross-modal attention over the egocentric observations of the agent. In contrast to other works, our key insight is that the association between language and vision is stronger when it occurs in explicit spatial representations. In this work, we propose a cross-modal map learning model for vision-and-language navigation that first learns to predict the top-down semantics on an egocentric map for both observed and unobserved regions, and then predicts a path towards the goal as a set of waypoints. In both cases, the prediction is informed by the language through cross-modal attention mechanisms. We experimentally test the basic hypothesis that language-driven navigation can be solved given a map, and then show competitive results on the full VLN-CE benchmark.
****

Motion-Aware Contrastive Video Representation Learning via Foreground-Background Merging
Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, Hongkai Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9716-9726
In light of the success of contrastive learning in the image domain, current self-supervised video representation learning methods usually employ contrastive loss to facilitate video representation learning. When naively pulling two augmented views of a video closer, the model however tends to learn the common static background as a shortcut but fails to capture the motion information, a phenomenon dubbed as background bias. Such bias makes the model suffer from weak generalization ability, leading to worse performance on downstream tasks such as action recognition. To alleviate such bias, we propose Foreground-background Merging (FAME) to deliberately compose the moving foreground region of the selected video onto the static background of others. Specifically, without any off-the-shelf detector, we extract the moving foreground out of background regions via the frame difference and color statistics, and shuffle the background regions among the videos. By leveraging the semantic consistency between the original clips and the fused ones, the model focuses more on the motion patterns and is debiased from the background shortcut. Extensive experiments demonstrate that FAME can effectively resist background cheating and thus achieve the state-of-the-art performance on downstream tasks across UCF101, HMDB51, and Diving48 datasets. The code and configurations are released at https://github.com/Mark12Ding/FAME.
****

Incremental Transformer Structure Enhanced Image Inpainting With Masking Positional Encoding
Qiaole Dong, Chenjie Cao, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11358-11368
Image inpainting has made significant advances in recent years. However, it is still challenging to recover corrupted images with both vivid textures and reasonable structures. Some specific methods can only tackle regular textures while losing holistic structures due to the limited receptive fields of convolutional neural networks (CNNs). On the other hand, attention-based models can learn better long-range dependency for the structure recovery, but they are limited by the heavy computation for inference with large image sizes. To address these issues,

we propose to leverage an additional structure restorer to facilitate the image inpainting incrementally. The proposed model restores holistic image structures with a powerful attention-based transformer model in a fixed low-resolution sketch space. Such a grayscale space is easy to be upsampled to larger scales to convey correct structural information. Our structure restorer can be integrated with other pretrained inpainting models efficiently with the zero-initialized residual addition. Furthermore, a masking positional encoding strategy is utilized to improve the performance of the proposed model with large irregular masks. Extensive experiments on various datasets validate the efficacy of our model compared with other competitors.
****

Pointly-Supervised Instance Segmentation
Bowen Cheng, Omkar Parkhi, Alexander Kirillov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2617-2626
We propose an embarrassingly simple point annotation scheme to collect weak supervision for instance segmentation. In addition to bounding boxes, we collect binary labels for a set of points uniformly sampled inside each bounding box. We show that the existing instance segmentation models developed for full mask supervision can be seamlessly trained with point-based supervision collected via our scheme. Remarkably, Mask R-CNN trained on COCO, PASCAL VOC, Cityscapes, and LVIS with only 10 annotated random points per object achieves 94%-98% of its fully-supervised performance, setting a strong baseline for weakly-supervised instance segmentation. The new point annotation scheme is approximately 5 times faster than annotating full object masks, making high-quality instance segmentation more accessible in practice. Inspired by the point-based annotation form, we propose a modification to PointRend instance segmentation module. For each object, the new architecture, called Implicit PointRend, generates parameters for a function that makes the final point-level mask prediction. Implicit PointRend is more straightforward and uses a single point-level mask loss. Our experiments show that the new module is more suitable for the point-based supervision.
****

Cross-Modal Clinical Graph Transformer for Ophthalmic Report Generation
Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, Xiaojun Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20656-20665
Automatic generation of ophthalmic reports using data-driven neural networks has great potential in clinical practice. When writing a report, ophthalmologists make inferences with prior clinical knowledge. This knowledge has been neglected in prior medical report generation methods. To endow models with the capability of incorporating expert knowledge, we propose a Cross-modal clinical Graph Transformer (CGT) for ophthalmic report generation (ORG), in which clinical relation triples are injected into the visual features as prior knowledge to drive the decoding procedure. However, two major common Knowledge Noise (KN) issues may affect models' effectiveness. 1) Existing general biomedical knowledge bases such as the UMLS may not align meaningfully to the specific context and language of the report, limiting their utility for knowledge injection. 2) Incorporating too much knowledge may divert the visual features from their correct meaning. To overcome these limitations, we design an automatic information extraction scheme based on natural language processing to obtain clinical entities and relations directly from in-domain training reports. Given a set of ophthalmic images, our CGT first restores a sub-graph from the clinical graph and injects the restored triples into visual features. Then visible matrix is employed during the encoding procedure to limit the impact of knowledge. Finally, reports are predicted by the encoded cross-modal features via a Transformer decoder. Extensive experiments on the large-scale FFA-IR benchmark demonstrate that the proposed CGT is able to outperform previous benchmark methods and achieve state-of-the-art performances.
****

Human-Object Interaction Detection via Disentangled Transformer
Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn

ition (CVPR), 2022, pp. 19568-19577
Human-Object Interaction Detection tackles the problem of joint localization and classification of human object interactions. Existing HOI transformers either adopt a single decoder for triplet prediction, or utilize two parallel decoders to detect individual objects and interactions separately, and compose triplets by a matching process. In contrast, we decouple the triplet prediction into human-object pair detection and interaction classification. Our main motivation is that detecting the human-object instances and classifying interactions accurately needs to learn representations that focus on different regions. To this end, we present Disentangled Transformer, where both encoder and decoder are disentangled to facilitate learning of two subtasks. To associate the predictions of disentangled decoders, we first generate a unified representation for HOI triplets with a base decoder, and then utilize it as input feature of each disentangled decoder. Extensive experiments show that our method outperforms prior work on two public HOI benchmarks by a sizeable margin. Code will be available.
****

DINE: Domain Adaptation From Single and Multiple Black-Box Predictors
Jian Liang, Dapeng Hu, Jiashi Feng, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8003-8013
To ease the burden of labeling, unsupervised domain adaptation (UDA) aims to transfer knowledge in previous and related labeled datasets (sources) to a new unlabeled dataset (target). Despite impressive progress, prior methods always need to access the raw source data and develop data-dependent alignment approaches to recognize the target samples in a transductive learning manner, which may raise privacy concerns from source individuals. Several recent studies resort to an alternative solution by exploiting the well-trained white-box model from the source domain, yet, it may still leak the raw data via generative adversarial learning. This paper studies a practical and interesting setting for UDA, where only black-box source models (i.e., only network predictions are available) are provided during adaptation in the target domain. To solve this problem, we propose a new two-step knowledge adaptation framework called DIstill and fine-tuNE (DINE). Taking into consideration the target data structure, DINE first distills the knowledge from the source predictor to a customized target model, then fine-tunes the distilled model to further fit the target domain. Besides, neural networks are not required to be identical across domains in DINE, even allowing effective adaptation on a low-resource device. Empirical results on three UDA scenarios (i.e., single-source, multi-source, and partial-set) confirm that DINE achieves highly competitive performance compared to state-of-the-art data-dependent approaches. Code is available at https://github.com/tim-learn/DINE/.
****

LGT-Net: Indoor Panoramic Room Layout Estimation With Geometry-Aware Transformer Network
Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, Ming Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1654-1663
3D room layout estimation by a single panorama using deep neural networks has made great progress. However, previous approaches can not obtain efficient geometry awareness of room layout with the only latitude of boundaries or horizon-depth. We present that using horizon-depth along with room height can obtain omnidirectional-geometry awareness of room layout in both horizontal and vertical directions. In addition, we propose a planar-geometry aware loss function with normals and gradients of normals to supervise the planeness of walls and turning of corners. We propose an efficient network, LGT-Net, for room layout estimation, which contains a novel Transformer architecture called SWG-Transformer to model geometry relations. SWG-Transformer consists of (Shifted) Window Blocks and Global Blocks to combine the local and global geometry relations. Moreover, we design a novel relative position embedding of Transformer to enhance the spatial identification ability for the panorama. Experiments show that the proposed LGT-Net achieves better performance than current state-of-the-arts (SOTA) on benchmark datasets.

****
CRIS: CLIP-Driven Referring Image Segmentation
Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, Tongliang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11686-11695
Referring image segmentation aims to segment a referent via a natural linguistic expression. Due to the distinct data properties between text and image, it is challenging for a network to well align text and pixel-level features. Existing approaches use pretrained models to facilitate learning, yet separately transfer the language/vision knowledge from pretrained models, ignoring the multi-modal corresponding information. Inspired by the recent advance in Contrastive Language-Image Pretraining (CLIP), in this paper, we propose an end-to-end CLIP-Driven Referring Image Segmentation framework (CRIS). To transfer the multi-modal knowledge effectively, CRIS resorts to vision-language decoding and contrastive learning for achieving the text-to-pixel alignment. More specifically, we design a vision-language decoder to propagate fine-grained semantic information from textual representations to each pixel-level activation, which promotes consistency between the two modalities. In addition, we present text-to-pixel contrastive learning to explicitly enforce the text feature similar to the related pixel-level features and dissimilar to the irrelevances. The experimental results on three benchmark datasets demonstrate that our proposed framework significantly outperforms the state-of-the-art performance without any post-processing.
****
Multi-View Mesh Reconstruction With Neural Deferred Shading
Markus Worchel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, Peter Eisert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6187-6197
We propose an analysis-by-synthesis method for fast multi-view 3D reconstruction of opaque objects with arbitrary materials and illumination. State-of-the-art methods use both neural surface representations and neural rendering. While flexible, neural surface representations are a significant bottleneck in optimization runtime. Instead, we represent surfaces as triangle meshes and build a differentiable rendering pipeline around triangle rasterization and neural shading. The renderer is used in a gradient descent optimization where both a triangle mesh and a neural shader are jointly optimized to reproduce the multi-view images. We evaluate our method on a public 3D reconstruction dataset and show that it can match the reconstruction accuracy of traditional baselines and neural approaches while surpassing them in optimization runtime. Additionally, we investigate the shader and find that it learns an interpretable representation of appearance, enabling applications such as 3D material editing.
****
CVF-SID: Cyclic Multi-Variate Function for Self-Supervised Image Denoising by Disentangling Noise From Image
Reyhaneh Neshatavar, Mohsen Yavartanoo, Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17583-17591
Recently, significant progress has been made on image denoising with strong supervision from large-scale datasets. However, obtaining well-aligned noisy-clean training image pairs for each specific scenario is complicated and costly in practice. Consequently, applying a conventional supervised denoising network on in-the-wild noisy inputs is not straightforward. Although several studies have challenged this problem without strong supervision, they rely on less practical assumptions and cannot be applied to practical situations directly. To address the aforementioned challenges, we propose a novel and powerful self-supervised denoising method called CVF-SID based on a Cyclic multi-Variate Function (CVF) module and a self-supervised image disentangling (SID) framework. The CVF module can output multiple decomposed variables of the input and take a combination of the outputs back as an input in a cyclic manner. Our CVF-SID can disentangle a clean image and noise maps from the input by leveraging various self-supervised loss terms. Unlike several methods that only consider the signal-independent noise model

s, we also deal with signal-dependent noise components for real-world applicatio
ns. Furthermore, we do not rely on any prior assumptions about the underlying no
ise distribution, making CVF-SID more generalizable toward realistic noise. Exte
nsive experiments on real-world datasets show that CVF-SID achieves state-of-the
-art self-supervised image denoising performance and is comparable to other exis
ting approaches. The code is publicly available from this link.
****

Infrared Invisible Clothing: Hiding From Infrared Detectors at Multiple Angles i
n Real World
Xiaopei Zhu, Zhanhao Hu, Siyuan Huang, Jianmin Li, Xiaolin Hu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p
p. 13317-13326
Thermal infrared imaging is widely used in body temperature measurement, securit
y monitoring, and so on, but its safety research attracted attention only in rec
ent years. We proposed the infrared adversarial clothing, which could fool infra
red pedestrian detectors at different angles. We simulated the process from clot
h to clothing in the digital world and then designed the adversarial "QR code" p
attern. The core of our method is to design a basic pattern that can be expanded
 periodically, and make the pattern after random cropping and deformation still
have an adversarial effect, then we can process the flat cloth with an adversari
al pattern into any 3D clothes. The results showed that the optimized "QR code"
pattern lowered the Average Precision (AP) of YOLOv3 by 87.7%, while the random
"QR code" pattern and blank pattern lowered the AP of YOLOv3 by 57.9% and 30.1%,
 respectively, in the digital world. We then manufactured an adversarial shirt w
ith a new material: aerogel. Physical-world experiments showed that the adversar
ial "QR code" pattern clothing lowered the AP of YOLOv3 by 64.6%, while the rand
om "QR code" pattern clothing and fully heat-insulated clothing lowered the AP o
f YOLOv3 by 28.3% and 22.8%, respectively. We used the model ensemble technique
to improve the attack transferability to unseen models.
****

Distribution-Aware Single-Stage Models for Multi-Person 3D Pose Estimation
Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, Si Liu; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 13096-13105
In this paper, we present a novel Distribution-Aware Single-stage (DAS) model fo
r tackling the challenging multi-person 3D pose estimation problem. Different fr
om existing top-down and bottom-up methods, the proposed DAS model simultaneousl
y localizes person positions and their corresponding body joints in the 3D camer
a space in a one-pass manner. This leads to a simplified pipeline with enhanced
efficiency. In addition, DAS learns the true distribution of body joints for the
 regression of their positions, rather than making a simple Laplacian or Gaussia
n assumption as previous works. This provides valuable priors for model predicti
on and thus boosts the regression-based scheme to achieve competitive performanc
e with volumetric-base ones. Moreover, DAS exploits a recursive update strategy
for progressively approaching to regression target, alleviating the optimization
 difficulty and further lifting the regression performance. DAS is implemented w
ith a fully Convolutional Neural Network and end-to-end learnable. Comprehensive
 experiments on benchmarks CMU Panoptic and MuPoTS-3D demonstrate the superior e
fficiency of the proposed DAS model, specifically 1.5x speedup over previous bes
t model, and its stat-of-the-art accuracy for multi-person 3D pose estimation.
****

FaceFormer: Speech-Driven 3D Facial Animation With Transformers
Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, Taku Komura; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 18770-18780
Speech-driven 3D facial animation is challenging due to the complex geometry of
human faces and the limited availability of 3D audio-visual data. Prior works ty
pically focus on learning phoneme-level features of short audio windows with lim
ited context, occasionally resulting in inaccurate lip movements. To tackle this
 limitation, we propose a Transformer-based autoregressive model, FaceFormer, wh

ich encodes the long-term audio context and autoregressively predicts a sequence of animated 3D face meshes. To cope with the data scarcity issue, we integrate the self-supervised pre-trained speech representations. Also, we devise two biased attention mechanisms well suited to this specific task, including the biased cross-modal multi-head (MH) attention and the biased causal MH self-attention with a periodic positional encoding strategy. The former effectively aligns the audio-motion modalities, whereas the latter offers abilities to generalize to longer audio sequences. Extensive experiments and a perceptual user study show that our approach outperforms the existing state-of-the-arts. The code and the video are available at: https://evelynfan.github.io/audio2face/.

****

Exploring Patch-Wise Semantic Relation for Contrastive Learning in Image-to-Image Translation Tasks

Chanyong Jung, Gihyun Kwon, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18260-18269

Recently, contrastive learning-based image translation methods have been proposed, which contrasts different spatial locations to enhance the spatial correspondence. However, the methods often ignore the diverse semantic relation within the images. To address this, here we propose a novel semantic relation consistency (SRC) regularization along with the decoupled contrastive learning (DCL), which utilize the diverse semantics by focusing on the heterogeneous semantics between the image patches of a single image. To further improve the performance, we present a hard negative mining by exploiting the semantic relation. We verified our method for three tasks: single-modal and multi-modal image translations, and GAN compression task for image translation. Experimental results confirmed the state-of-art performance of our method in all the three tasks.

****

High-Resolution Face Swapping via Latent Semantics Disentanglement

Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7642-7651

We present a novel high-resolution face swapping method using the inherent prior knowledge of a pre-trained GAN model. Although previous research can leverage generative priors to produce high-resolution results, their quality can suffer from the entangled semantics of the latent space. We explicitly disentangle the latent semantics by utilizing the progressive nature of the generator, deriving structure attributes from the shallow layers and appearance attributes from the deeper ones. Identity and pose information within the structure attributes are further separated by introducing a landmark-driven structure transfer latent direction. The disentangled latent code produces rich generative features that incorporate feature blending to produce a plausible swapping result. We further extend our method to video face swapping by enforcing two spatio-temporal constraints on the latent space and the image space. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art image/video face swapping methods in terms of hallucination quality and consistency.

****

Searching the Deployable Convolution Neural Networks for GPUs

Linnan Wang, Chenhan Yu, Satish Salian, Slawomir Kierat, Szymon Migacz, Alex Fit Florea; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12227-12236

Customizing Convolution Neural Networks (CNN) for production use has been a challenging task for DL practitioners. This paper intends to expedite the model customization with a model hub that contains the optimized models tiered by their inference latency using Neural Architecture Search (NAS). To achieve this goal, we build a distributed NAS system to search on a novel search space that consists of prominent factors to impact latency and accuracy. Since we target GPU, we name the NAS optimized models as GPUNet, which establishes a new SOTA Pareto frontier in inference latency and accuracy. Within 1ms, GPUNet is 2x faster than EfficientNet-X and FBNetV3 with even better accuracy. We also validate GPUNet on detection tasks, and GPUNet consistently outperforms EfficientNet-X and FBNetV3 on C

OCO detection tasks in both latency and accuracy. All of these data validate tha t our NAS system is effective and generic to handle different design tasks. With this NAS system, we expand GPUNet to cover more latency groups to be directly r eusable to DL practitioners in various deployment scenarios.
****

Sparse Local Patch Transformer for Robust Face Alignment and Landmarks Inherent Relation Learning
Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, Min Xu; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4052-4061
Heatmap regression methods have dominated face alignment area in recent years wh ile they ignore the inherent relation between different landmarks. In this paper , we propose a Sparse Local Patch Transformer (SLPT) for learning the inherent r elation. The SLPT generates the representation of each single landmark from a lo cal patch and aggregates them by an adaptive inherent relation based on the atte ntion mechanism. The subpixel coordinate of each landmark is predicted independe ntly based on the aggregated feature. Moreover, a coarse-to-fine framework is fu rther introduced to incorporate with the SLPT, which enables the initial landmar ks to gradually converge to the target facial landmarks using fine-grained featu res from dynamically resized local patches. Extensive experiments carried out on three popular benchmarks, including WFLW, 300W and COFW, demonstrate that the p roposed method works at the state-of-the-art level with much less computational complexity by learning the inherent relation between facial landmarks. The code is available at the project website.
****

DeepFake Disrupter: The Detector of DeepFake Is My Friend
Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, Chang Xu; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14 920-14929
In recent years, with the advances of generative models, many powerful face mani pulation systems have been developed based on Deep Neural Networks (DNNs), calle d DeepFakes. If DeepFakes are not controlled timely and properly, they would bec ome a real threat to both celebrities and ordinary people. Precautions such as a dding perturbations to the source inputs will make DeepFake results look distort ed from the perspective of human eyes. However, previous method doesn't explore whether the disrupted images can still spoof DeepFake detectors. This is critica l for many applications where DeepFake detectors are used to discriminate betwee n DeepFake data and real data due to the huge cost of examining a large amount o f data manually. We argue that the detectors do not share a similar perspective as human eyes, which might still be spoofed by the disrupted data. Besides, the existing disruption methods rely on iteration-based perturbation generation algo rithms, which is time-consuming. In this paper, we propose a novel DeepFake disr uption algorithm called "DeepFake Disrupter". By training a perturbation generat or, we can add the human-imperceptible perturbations to source images that need to be protected without any backpropagation update. The DeepFake results of thes e protected source inputs would not only look unrealistic by the human eye but a lso can be distinguished by DeepFake detectors easily. For example, experimental results show that by adding our trained perturbations, fake images generated by StarGAN can result in a 10 20% increase in F1-score evaluated by various Deep Fake detectors.
****

Rotationally Equivariant 3D Object Detection
Hong-Xing Yu, Jiajun Wu, Li Yi; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2022, pp. 1456-1464
Rotation equivariance has recently become a strongly desired property in the 3D deep learning community. Yet most existing methods focus on equivariance regardi ng a global input rotation while ignoring the fact that rotation symmetry has it s own spatial support. Specifically, we consider the object detection problem in 3D scenes, where an object bounding box should be equivariant regarding the obj ect pose, independent of the scene motion. This suggests a new desired property

we call object-level rotation equivariance. To incorporate object-level rotation equivariance into 3D object detectors, we need a mechanism to extract equivariant features with local object-level spatial support while being able to model cross-object context information. To this end, we propose Equivariant Object detection Network (EON) with a rotation equivariance suspension design to achieve object-level equivariance. EON can be applied to modern point cloud object detectors, such as VoteNet and PointRCNN, enabling them to exploit object rotation symmetry in scene-scale inputs. Our experiments on both indoor scene and autonomous driving datasets show that significant improvements are obtained by plugging our EON design into existing state-of-the-art 3D object detectors. Project website: https://kovenyu.com/EON/.

****

Accelerating DETR Convergence via Semantic-Aligned Matching

Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 949-958

The recently developed DEtection TRansformer (DETR) establishes a new object detection paradigm by eliminating a series of hand-crafted components. However, DETR suffers from extremely slow convergence, which increases the training cost significantly. We observe that the slow convergence is largely attributed to the complication in matching object queries with target features in different feature embedding spaces. This paper presents SAM-DETR, a Semantic-Aligned-Matching DETR that greatly accelerates DETR's convergence without sacrificing its accuracy. SAM-DETR addresses the convergence issue from two perspectives. First, it projects object queries into the same embedding space as encoded image features, where the matching can be accomplished efficiently with aligned semantics. Second, it explicitly searches salient points with the most discriminative features for semantic-aligned matching, which further speeds up the convergence and boosts detection accuracy as well. Being like a plug and play, SAM-DETR complements existing convergence solutions well yet only introduces slight computational overhead. Extensive experiments show that the proposed SAM-DETR achieves superior convergence as well as competitive detection accuracy. The implementation codes are publicly available at https://github.com/ZhangGongjie/SAM-DETR.

****

Long-Short Temporal Contrastive Learning of Video Transformers

Jue Wang, Gedas Bertasius, Du Tran, Lorenzo Torresani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14010-14020

Video transformers have recently emerged as a competitive alternative to 3D CNNs for video understanding. However, due to their large number of parameters and reduced inductive biases, these models require supervised pretraining on large-scale image datasets to achieve top performance. In this paper, we empirically demonstrate that self-supervised pretraining of video transformers on video-only datasets can lead to action recognition results that are on par or better than those obtained with supervised pretraining on large-scale image datasets, even massive ones such as ImageNet-21K. Since transformer-based models are effective at capturing dependencies over extended temporal spans, we propose a simple learning procedure that forces the model to match a long-term view to a short-term view of the same video. Our approach, named Long-Short Temporal Contrastive Learning (LSTCL), enables video transformers to learn an effective clip-level representation by predicting temporal context captured from a longer temporal extent. To demonstrate the generality of our findings, we implement and validate our approach under three different self-supervised contrastive learning frameworks (MoCo v3, BYOL, SimSiam) using two distinct video-transformer architectures, including an improved variant of the Swin Transformer augmented with space-time attention. We conduct a thorough ablation study and show that LSTCL achieves competitive performance on multiple video benchmarks and represents a convincing alternative to supervised image-based pretraining.

****

Vision Transformer With Deformable Attention

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4794-4803

Transformers have recently shown superior performances on various vision tasks. The large, sometimes even global, receptive field endows Transformer models with higher representation power over their CNN counterparts. Nevertheless, simply enlarging receptive field also gives rise to several concerns. On the one hand, using dense attention e.g., in ViT, leads to excessive memory and computational cost, and features can be influenced by irrelevant parts which are beyond the region of interests. On the other hand, the sparse attention adopted in PVT or Swin Trans-former is data agnostic and may limit the ability to model long range relations. To mitigate these issues, we propose a novel deformable self-attention module, where the positions of key and value pairs in self-attention are selected in a data-dependent way. This flexible scheme enables the self-attention module to focus on relevant regions and cap-ture more informative features. On this basis, we present Deformable Attention Transformer, a general backbone model with deformable attention for both image classifi-cation and dense prediction tasks. Extensive experiments show that our models achieve consistently improved results on comprehensive benchmarks.

****

Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture

Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, Derek Hoiem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16399-16409

Computer vision systems today are primarily N-purpose systems, designed and trained for a predefined set of tasks. Adapting such systems to new tasks is challenging and often requires non-trivial modifications to the network architecture (e.g. adding new output heads) or training process (e.g. adding new losses). To reduce the time and expertise required to develop new applications, we would like to create general purpose vision systems that can learn and perform a range of tasks without any modification to the architecture or learning process. In this paper, we propose GPV-1, a task-agnostic vision-language architecture that can learn and perform tasks that involve receiving an image and producing text and/or bounding boxes, including classification, localization, visual question answering, captioning, and more. We also propose evaluations of generality of architecture, skill-concept transfer, and learning efficiency that may inform future work on general purpose vision. Our experiments indicate GPV-1 is effective at multiple tasks, reuses some concept knowledge across tasks, can perform the Referring Expressions task zero-shot, and further improves upon the zero-shot performance using a few training samples.

****

Deep Vanishing Point Detection: Geometric Priors Make Dataset Variations Vanish

Yancong Lin, Ruben Wiersma, Silvia L. Pintea, Klaus Hildebrandt, Elmar Eisemann, Jan C. van Gemert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6103-6113

Deep learning has improved vanishing point detection in images. Yet, deep networks require expensive annotated datasets trained on costly hardware and do not generalize to even slightly different domains, and minor problem variants. Here, we address these issues by injecting deep vanishing point detection networks with prior knowledge. This prior knowledge no longer needs to be learned from data, saving valuable annotation efforts and compute, unlocking realistic few-sample scenarios, and reducing the impact of domain changes. Moreover, the interpretability of the priors allows to adapt deep networks to minor problem variations such as switching between Manhattan and non-Manhattan worlds. We seamlessly incorporate two geometric priors: (i) Hough Transform -- mapping image pixels to straight lines, and (ii) Gaussian sphere -- mapping lines to great circles whose intersections denote vanishing points. Experimentally, we ablate our choices and show comparable accuracy to existing models in the large-data setting. We validate our model's improved data efficiency, robustness to domain changes, adaptability t

o non-Manhattan settings.
****
RM-Depth: Unsupervised Learning of Recurrent Monocular Depth in Dynamic Scenes
Tak-Wai Hui; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 1675-1684
Unsupervised methods have showed promising results on monocular depth estimation . However, the training data must be captured in scenes without moving objects. To push the envelope of accuracy, recent methods tend to increase their model pa rameters. In this paper, an unsupervised learning framework is proposed to joint ly predict monocular depth and complete 3D motion including the motions of movin g objects and camera. (1) Recurrent modulation units are used to adaptively and iteratively fuse encoder and decoder features. This improves the single-image de pth inference without overspending model parameters. (2) Instead of using a sing le set of filters for upsampling, multiple sets of filters are devised for the r esidual upsampling. This facilitates the learning of edge-preserving filters and leads to the improved performance. (3) A warping-based network is used to estim ate a motion field of moving objects without using semantic priors. This breaks down the requirement of scene rigidity and allows to use general videos for the unsupervised learning. The motion field is further regularized by an outlier-awa re training loss. Despite the depth model just uses a single image in test time and 2.97M parameters, it achieves state-of-the-art results on the KITTI and City scapes benchmarks.
****
LiT: Zero-Shot Transfer With Locked-Image Text Tuning
Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexand er Kolesnikov, Lucas Beyer; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 18123-18133
This paper presents contrastive-tuning, a simple method employing contrastive tr aining to align image and text models while still taking advantage of their pre- training. In our empirical study we find that locked pre-trained image models wi th unlocked text models work best. We call this instance of contrastive-tuning " Locked-image Tuning" (LiT), which just teaches a text model to read out good rep resentations from a pre-trained image model for new tasks. A LiT model gains the capability of zero-shot transfer to new vision tasks, such as image classificat ion or retrieval. The proposed LiT is widely applicable; it works reliably with multiple pre-training methods (supervised and unsupervised) and across diverse a rchitectures (ResNet, Vision Transformers and MLP-Mixer) using three different i mage-text datasets. With the transformer-based pre-trained ViT-g/14 model, the L iT model achieves 84.5% zero-shot transfer accuracy on the ImageNet test set, an d 81.1% on the challenging out-of-distribution ObjectNet test set.
****
Cloning Outfits From Real-World Images to 3D Characters for Generalizable Person Re-Identification
Yanan Wang, Xuezhi Liang, Shengcai Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4900-4909
Recently, large-scale synthetic datasets are shown to be very useful for general izable person re-identification. However, synthesized persons in existing datase ts are mostly cartoon-like and in random dress collocation, which limits their p erformance. To address this, in this work, an automatic approach is proposed to directly clone the whole outfits from real-world person images to virtual 3D cha racters, such that any virtual person thus created will appear very similar to i ts real-world counterpart. Specifically, based on UV texture mapping, two clonin g methods are designed, namely registered clothes mapping and homogeneous cloth expansion. Given clothes keypoints detected on person images and labeled on regu lar UV maps with clear clothes structures, registered mapping applies perspectiv e homography to warp real-world clothes to the counterparts on the UV map. As fo r invisible clothes parts and irregular UV maps, homogeneous expansion segments a homogeneous area on clothes as a realistic cloth pattern or cell, and expand t he cell to fill the UV map. Furthermore, a similarity-diversity expansion strate gy is proposed, by clustering person images, sampling images per cluster, and cl

oning outfits for 3D character generation. This way, virtual persons can be scaled up densely in visual similarity to challenge model learning, and diversely in population to enrich sample distribution. Finally, by rendering the cloned characters in Unity3D scenes, a more realistic virtual dataset called ClonedPerson is created, with 5,621 identities and 887,766 images. Experimental results show that the model trained on ClonedPerson has a better generalization performance, superior to that trained on other popular real-world and synthetic person re-identification datasets. The ClonedPerson project is available at https://github.com/Yanan-Wang-cs/ClonedPerson.

****

GeoNeRF: Generalizing NeRF With Geometry Priors

Mohammad Mahdi Johari, Yann Lepoittevin, François Fleuret; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18365-18375

We present GeoNeRF, a generalizable photorealistic novel view synthesis method based on neural radiance fields. Our approach consists of two main stages: a geometry reasoner and a renderer. To render a novel view, the geometry reasoner first constructs cascaded cost volumes for each nearby source view. Then, using a Transformer-based attention mechanism and the cascaded cost volumes, the renderer infers geometry and appearance, and renders detailed images via classical volume rendering techniques. This architecture, in particular, allows sophisticated occlusion reasoning, gathering information from consistent source views. Moreover, our method can easily be fine-tuned on a single scene, and renders competitive results with per-scene optimized neural rendering methods with a fraction of computational cost. Experiments show that GeoNeRF outperforms state-of-the-art generalizable neural rendering models on various synthetic and real datasets. Lastly, with a slight modification to the geometry reasoner, we also propose an alternative model that adapts to RGBD images. This model directly exploits the depth information often available thanks to depth sensors. The implementation code is available at https://www.idiap.ch/paper/geonerf.

****

ABPN: Adaptive Blend Pyramid Network for Real-Time Local Retouching of Ultra High-Resolution Photo

Biwen Lei, Xiefan Guo, Hongyu Yang, Miaomiao Cui, Xuansong Xie, Di Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2108-2117

Photo retouching finds many applications in various fields. However, most existing methods are designed for global retouching and seldom pay attention to the local region, while the latter is actually much more tedious and time-consuming in photography pipelines. In this paper, we propose a novel adaptive blend pyramid network, which aims to achieve fast local retouching on ultra high-resolution photos. The network is mainly composed of two components: a context-aware local retouching layer (LRL) and an adaptive blend pyramid layer (BPL). The LRL is designed to implement local retouching on low-resolution images, giving full consideration of the global context and local texture information, and the BPL is then developed to progressively expand the low-resolution results to the higher ones, with the help of the proposed adaptive blend module and refining module. Our method outperforms the existing methods by a large margin on two local photo retouching tasks and exhibits excellent performance in terms of running speed, achieving real-time inference on 4K images with a single NVIDIA Tesla P100 GPU. Moreover, we introduce the first high-definition cloth retouching dataset CRHD-3K to promote the research on local photo retouching. The dataset is available at https://github.com/youngLBW/CRHD-3K.

****

PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation With Photometrically Challenging Objects

Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, Benjamin Busam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21222-21231

Object pose estimation is crucial for robotic applications and augmented reality. Beyond instance level 6D object pose estimation methods, estimating category-level pose and shape has become a promising trend. As such, a new research field needs to be supported by well-designed datasets. To provide a benchmark with high-quality ground truth annotations to the community, we introduce a multimodal dataset for category-level object pose estimation with photometrically challenging objects termed PhoCaL. PhoCaL comprises 60 high quality 3D models of household objects over 8 categories including highly reflective, transparent and symmetric objects. We developed a novel robot-supported multi-modal (RGB, depth, polarisation) data acquisition and annotation process. It ensures sub-millimeter accuracy of the pose for opaque textured, shiny and transparent objects, no motion blur and perfect camera synchronisation. To set a benchmark for our dataset, state-of-the-art RGB-D and monocular RGB methods are evaluated on the challenging scenes of PhoCaL.

****

Neural Compression-Based Feature Learning for Video Restoration

Cong Huang, Jiahao Li, Bin Li, Dong Liu, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5872-5881

Most existing deep learning (DL)-based video restoration methods focus on the network structure design to better extract temporal features but ignore how to utilize these extracted temporal features efficiently. The temporal features usually contain various noisy and irrelative information, and they may interfere with the restoration of the current frame. This paper proposes learning noise-robust feature representations to help video restoration. From information theory, we know the noisy data generally has a high degree of uncertainty, thus we design a neural compression module to filter the noise with large uncertainty and refine the features. Our compression module adopts a spatial-channel-wise quantization mechanism to adaptively filter the noise and purify the features with different content characteristics to achieve robustness to noise. The information entropy loss is used to guide the learning of the compression module and helps it preserve the most useful information. Experiments show that our method can significantly boost the performance on video denoising. Under noise level 50, we obtain 0.13 dB improvement over BasicVSR++ with only 0.23x FLOPs. Meanwhile, our method also achieves SOTA results on video deraining and dehazing.

****

Expanding Low-Density Latent Regions for Open-Set Object Detection

Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9591-9600

Modern object detectors have achieved impressive progress under the close-set setup. However, open-set object detection (OSOD) remains challenging since objects of unknown categories are often misclassified to existing known classes. In this work, we propose to identify unknown objects by separating high/low-density regions in the latent space, based on the consensus that unknown objects are usually distributed in low-density latent regions. As traditional threshold-based methods only maintain limited low-density regions, which cannot cover all unknown objects, we present a novel Open-set Detector (OpenDet) with expanded low-density regions. To this aim, we equip OpenDet with two learners, Contrastive Feature Learner (CFL) and Unknown Probability Learner (UPL). CFL performs instance-level contrastive learning to encourage compact features of known classes, leaving more low-density regions for unknown classes; UPL optimizes unknown probability based on the uncertainty of predictions, which further divides more low-density regions around the cluster of known classes. Thus, unknown objects in low-density regions can be easily identified with the learned unknown probability. Extensive experiments demonstrate that our method can significantly improve the OSOD performance, e.g., OpenDet reduces the Absolute Open-Set Errors by 25%-35% on six OSOD benchmarks. Code is available at: https://github.com/csuhan/opendet2.

****

Drop the GAN: In Defense of Patches Nearest Neighbors As Single Image Generative Models

Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, Michal Irani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13460-13469

Image manipulation dates back long before the deep learning era. The classical prevailing approaches were based on maximizing patch similarity between the input and generated output. Recently, single-image GANs were introduced as a superior and more sophisticated solution to image manipulation tasks. Moreover, they offered the opportunity not only to manipulate a given image, but also to generate a large and diverse set of different outputs from a single natural image. This gave rise to new tasks, which are considered "DL-only". However, despite their impressiveness, single-image GANs require long training time (usually hours) for each image and each task and often suffer from visual artifacts. In this paper we revisit the classical patch-based methods, and show that - unlike previously believed -- classical methods can be adapted to tackle these novel "GAN-only" tasks. Moreover, they do so better and faster than single-image GAN-based methods. More specifically, we show that: (i) by introducing slight modifications, classical patch-based methods are able to unconditionally generate diverse images based on a single natural image; (ii) the generated output visual quality exceeds that of single-image GANs by a large margin (confirmed both quantitatively and qualitatively); (iii) they are orders of magnitude faster (runtime reduced from hours to seconds).

****

Uformer: A General U-Shaped Transformer for Image Restoration

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17683-17693

In this paper, we present Uformer, an effective and efficient Transformer-based architecture for image restoration, in which we build a hierarchical encoder-decoder network using the Transformer block. In Uformer, there are two core designs. First, we introduce a novel locally-enhanced window (LeWin) Transformer block, which performs non-overlapping window-based self-attention instead of global self-attention. It significantly reduces the computational complexity on high resolution feature map while capturing local context. Second, we propose a learnable multi-scale restoration modulator in the form of a multi-scale spatial bias to adjust features in multiple layers of the Uformer decoder. Our modulator demonstrates superior capability for restoring details for various image restoration tasks while introducing marginal extra parameters and computational cost. Powered by these two designs, Uformer enjoys a high capability for capturing both local and global dependencies for image restoration. To evaluate our approach, extensive experiments are conducted on several image restoration tasks, including image denoising, motion deblurring, defocus deblurring and deraining. Without bells and whistles, our Uformer achieves superior or comparable performance compared with the state-of-the-art algorithms. The code and models are available at https://github.com/ZhendongWang6/Uformer.

****

Exploring Dual-Task Correlation for Pose Guided Person Image Generation

Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, Xiaohua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7713-7722

Pose Guided Person Image Generation (PGPIG) is the task of transforming a person image from the source pose to a given target pose. Most of the existing methods only focus on the ill-posed source-to-target task and fail to capture reasonable texture mapping. To address this problem, we propose a novel Dual-task Pose Transformer Network (DPTN), which introduces an auxiliary task (i.e., source-to-source task) and exploits the dual-task correlation to promote the performance of PGPIG. The DPTN is of a Siamese structure, containing a source-to-source self-reconstruction branch, and a transformation branch for source-to-target generation. By sharing partial weights between them, the knowledge learned by the source-to-source task can effectively assist the source-to-target learning. Furthermore, we bridge the two branches with a proposed Pose Transformer Module (PTM) to adap

tively explore the correlation between features from dual tasks. Such correlatio n can establish the fine-grained mapping of all the pixels between the sources a nd the targets, and promote the source texture transmission to enhance the detai ls of the generated target images. Extensive experiments show that our DPTN outp erforms state-of-the-arts in terms of both PSNR and LPIPS. In addition, our DPTN only contains 9.79 million parameters, which is significantly smaller than othe r approaches. Our code is available at: https://github.com/PangzeCheung/Dual-tas k-Pose-Transformer-Network.

****

Portrait Eyeglasses and Shadow Removal by Leveraging 3D Synthetic Data

Junfeng Lyu, Zhibo Wang, Feng Xu; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2022, pp. 3429-3439

In portraits, eyeglasses may occlude facial regions and generate cast shadows on faces, which degrades the performance of many techniques like face verification and expression recognition. Portrait eyeglasses removal is critical in handling these problems. However, completely removing the eyeglasses is challenging beca use the lighting effects (e.g., cast shadows) caused by them are often complex. In this paper, we propose a novel framework to remove eyeglasses as well as thei r cast shadows from face images. The method works in a detect-then-remove manner , in which eyeglasses and cast shadows are both detected and then removed from i mages. Due to the lack of paired data for supervised training, we present a new synthetic portrait dataset with both intermediate and final supervisions for bot h the detection and removal tasks. Furthermore, we apply a cross-domain techniqu e to fill the gap between the synthetic and real data. To the best of our knowle dge, the proposed technique is the first to remove eyeglasses and their cast sha dows simultaneously. The code and synthetic dataset are available at https://git hub.com/StoryMY/take-off-eyeglasses.

****

Neural Rays for Occlusion-Aware Image-Based Rendering

Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 7824-7833

We present a new neural representation, called Neural Ray (NeuRay), for the nove l view synthesis task. Recent works construct radiance fields from image feature s of input views to render novel view images, which enables the generalization t o new scenes. However, due to occlusions, a 3D point may be invisible to some in put views. On such a 3D point, these generalization methods will include inconsi stent image features from invisible views, which interfere with the radiance fie ld construction. To solve this problem, we predict the visibility of 3D points t o input views within our NeuRay representation. This visibility enables the radi ance field construction to focus on visible image features, which significantly improves its rendering quality. Meanwhile, a novel consistency loss is proposed to refine the visibility in NeuRay when finetuning on a specific scene. Experime nts demonstrate that our approach achieves state-of-the-art performance on the n ovel view synthesis task when generalizing to unseen scenes and outperforms per-scene optimization methods after finetuning.

****

Modeling 3D Layout for Group Re-Identification

Quan Zhang, Kaiheng Dang, Jian-Huang Lai, Zhanxiang Feng, Xiaohua Xie; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7512-7520

Group re-identification (GReID) attempts to correctly associate groups with the same members under different cameras. The main challenge is how to resist the me mbership and layout variations. Existing works attempt to incorporate layout mod eling on the basis of appearance features to achieve robust group representation s. However, layout ambiguity is introduced because these methods only consider t he 2D layout on the imaging plane. In this paper, we overcome the above limitati ons by 3D layout modeling. Specifically, we propose a novel 3D transformer (3DT) that reconstructs the relative 3D layout relationship among members, then appli es sampling and quantification to preset a series of layout tokens along three d

imensions, and selects the corresponding tokens as layout features for each memb
er. Furthermore, we build a synthetic GReID dataset, City1M, including 1.84M ima
ges, 45K persons and 11.5K groups with 3D annotations to alleviate data shortage
s and poor annotations. To the best of our knowledge, 3DT is the first work to a
ddress GReID with 3D perspective, and the City1M is the currently largest datase
t. Several experiments show the superiority of our 3DT and City1M. Our project h
as been released on https://github.com/LinlyAC/City1M-dataset.
****

Open-World Instance Segmentation: Exploiting Pseudo Ground Truth From Learned Pa
irwise Affinity
Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, Du Tran; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p
p. 4422-4432
Open-world instance segmentation is the task of grouping pixels into object inst
ances without any pre-determined taxonomy. This is challenging, as state-of-the-
art methods rely on explicit class semantics obtained from large labeled dataset
s, and out-of-domain evaluation performance drops significantly. Here we propose
 a novel approach for mask proposals, Generic Grouping Networks (GGNs), construc
ted without semantic supervision. Our approach combines a local measure of pixel
 affinity with instance-level mask supervision, producing a training regimen des
igned to make the model as generic as the data diversity allows. We introduce a
method for predicting Pairwise Affinities (PA), a learned local relationship bet
ween pairs of pixels. PA generalizes very well to unseen categories. From PA we
construct a large set of pseudo-ground-truth instance masks; combined with human
-annotated instance masks we train GGNs and significantly outperform the SOTA on
 open-world instance segmentation on various benchmarks including COCO, LVIS, AD
E20K, and UVO.
****

SIOD: Single Instance Annotated per Category per Image for Object Detection
Hanjun Li, Xingjia Pan, Ke Yan, Fan Tang, Wei-Shi Zheng; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 141
97-14206
Object detection under imperfect data receives great attention recently. Weakly
supervised object detection (WSOD) suffers from severe localization issues due t
o the lack of instance-level annotation, while semi-supervised object detection
(SSOD) remains challenging led by the inter-image discrepancy between labeled an
d unlabeled data. In this study, we propose the Single Instance annotated Object
 Detection (SIOD), requiring only one instance annotation for each existing cate
gory in an image. Degraded from inter-task (WSOD) or inter-image (SSOD) discrepa
ncies to the intra-image discrepancy, SIOD provides more reliable and rich prior
 knowledge for mining the rest of unlabeled instances and trades off the annotat
ion cost and performance. Under the SIOD setting, we propose a simple yet effect
ive framework, termed Dual-Mining (DMiner), which consists of a Similarity-based
 Pseudo Label Generating module (SPLG) and a Pixel-level Group Contrastive Learn
ing module (PGCL). SPLG firstly mines latent instances from feature representati
on space to alleviate the annotation missing problem. To avoid being misled by i
naccurate pseudo labels, we propose PGCL to boost the tolerance to false pseudo
labels. Extensive experiments on MS COCO verify the feasibility of the SIOD sett
ing and the superiority of the proposed method, which obtains consistent and sig
nificant improvements compared to baseline methods and achieves comparable resul
ts with fully supervised object detection (FSOD) methods with only 40% instances
 annotated.
****

Toward Fast, Flexible, and Robust Low-Light Image Enhancement
Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, Zhongxuan Luo; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 563
7-5646
Existing low-light image enhancement techniques are mostly not only difficult to
 deal with both visual quality and computational efficiency but also commonly in
valid in unknown complex scenarios. In this paper, we develop a new Self-Calibra

ted Illumination (SCI) learning framework for fast, flexible, and robust brighte
ning images in real-world low-light scenarios. To be specific, we establish a ca
scaded illumination learning process with weight sharing to handle this task. Co
nsidering the computational burden of the cascaded pattern, we construct the sel
f-calibrated module which realizes the convergence between results of each stage
, producing the gains that only use the single basic block for inference (yet ha
s not been exploited in previous works), which drastically diminishes computatio
n cost. We then define the unsupervised training loss to elevate the model capab
ility that can adapt general scenes. Further, we make comprehensive explorations
 to excavate SCI's inherent properties (lacking in existing works) including ope
ration-insensitive adaptability (acquiring stable performance under the settings
 of different simple operations) and model-irrelevant generality (can be applied
 to illumination-based existing works to improve performance). Finally, plenty o
f experiments and ablation studies fully indicate our superiority in both qualit
y and efficiency. Applications on low-light face detection and nighttime semanti
c segmentation fully reveal the latent practical values for SCI. The source code
 is available at https://github.com/vis-opt-group/SCI.
****

Online Learning of Reusable Abstract Models for Object Goal Navigation
Tommaso Campari, Leonardo Lamanna, Paolo Traverso, Luciano Serafini, Lamberto Ba
llan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2022, pp. 14870-14879
In this paper, we present a novel approach to incrementally learn an Abstract Mo
del of an unknown environment, and show how an agent can reuse the learned model
 for tackling the Object Goal Navigation task. The Abstract Model is a finite st
ate machine in which each state is an abstraction of a state of the environment,
 as perceived by the agent in a certain position and orientation. The perception
s are high-dimensional sensory data (e.g., RGB-D images), and the abstraction is
 reached by exploiting image segmentation and the Taskonomy model bank. The lear
ning of the Abstract Model is accomplished by executing actions, observing the r
eached state, and updating the Abstract Model with the acquired information. The
 learned models are memorized by the agent, and they are reused whenever it reco
gnizes to be in an environment that corresponds to the stored model. We investig
ate the effectiveness of the proposed approach for the Object Goal Navigation ta
sk, relying on public benchmarks. Our results show that the reuse of learned Abs
tract Models can boost performance on Object Goal Navigation.
****

Bridge-Prompt: Towards Ordinal Action Understanding in Instructional Videos
Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, Jiwen Lu;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2022, pp. 19880-19889
Action recognition models have shown a promising capability to classify human ac
tions in short video clips. In a real scenario, multiple correlated human action
s commonly occur in particular orders, forming semantically meaningful human act
ivities. Conventional action recognition approaches focus on analyzing single ac
tions. However, they fail to fully reason about the contextual relations between
 adjacent actions, which provide potential temporal logic for understanding long
 videos. In this paper, we propose a prompt-based framework, Bridge-Prompt (Br-P
rompt), to model the semantics across adjacent actions, so that it simultaneousl
y exploits both out-of-context and contextual information from a series of ordin
al actions in instructional videos. More specifically, we reformulate the indivi
dual action labels as integrated text prompts for supervision, which bridge the
gap between individual action semantics. The generated text prompts are paired w
ith corresponding video clips, and together co-train the text encoder and the vi
deo encoder via a contrastive approach. The learned vision encoder has a stronge
r capability for ordinal-action-related downstream tasks, e.g. action segmentati
on and human activity recognition. We evaluate the performances of our approach
on several video datasets: Georgia Tech Egocentric Activities (GTEA), 50Salads,
and the Breakfast dataset. Br-Prompt achieves state-of-the-art on multiple bench
marks. Code is available at: https://github.com/ttlmh/Bridge-Prompt.

****
SimMatch: Semi-Supervised Learning With Similarity Matching
Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14471-14481
Learning with few labeled data has been a longstanding problem in the computer vision and machine learning research community. In this paper, we introduced a new semi-supervised learning framework, SimMatch, which simultaneously considers semantic similarity and instance similarity. In SimMatch, the consistency regularization will be applied on both semantic-level and instance-level. The different augmented views of the same instance are encouraged to have the same class prediction and similar similarity relationship respected to other instances. Next, we instantiated a labeled memory buffer to fully leverage the ground truth labels on instance-level and bridge the gaps between the semantic and instance similarities. Finally, we proposed the unfolding and aggregation operation which allows these two similarities be isomorphically transformed with each other. In this way, the semantic and instance pseudo-labels can be mutually propagated to generate more high-quality and reliable matching targets. Extensive experimental results demonstrate that SimMatch improves the performance of semi-supervised learning tasks across different benchmark datasets and different settings. Notably, with 400 epochs of training, SimMatch achieves 67.2%, and 74.4% Top-1 Accuracy with 1% and 10% labeled examples on ImageNet, which significantly outperforms the baseline methods and is better than previous semi-supervised learning frameworks.
****
OrphicX: A Causality-Inspired Latent Variable Model for Interpreting Graph Neural Networks
Wanyu Lin, Hao Lan, Hao Wang, Baochun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13729-13738
This paper proposes a new eXplanation framework, called OrphicX, for generating causal explanations for any graph neural networks (GNNs) based on learned latent causal factors. Specifically, we construct a distinct generative model and design an objective function that encourages the generative model to produce causal, compact, and faithful explanations. This is achieved by isolating the causal factors in the latent space of graphs by maximizing the information flow measurements. We theoretically analyze the cause-effect relationships in the proposed causal graph, identify node attributes as confounders between graphs and GNN predictions, and circumvent such confounder effect by leveraging the backdoor adjustment formula. Our framework is compatible with any GNNs, and it does not require access to the process by which the target GNN produces its predictions. In addition, it does not rely on the linear-independence assumption of the explained features, nor require prior knowledge on the graph learning tasks. We show a proof-of-concept of OrphicX on canonical classification problems on graph data. In particular, we analyze the explanatory subgraphs obtained from explanations for molecular graphs (i.e., Mutag) and quantitatively evaluate the explanation performance with frequently occurring subgraph patterns. Empirically, we show that OrphicX can effectively identify the causal semantics for generating causal explanations, significantly outperforming its alternatives.
****
HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network
JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1496-1505
Hands are often severely occluded by objects, which makes 3D hand mesh estimation challenging. Previous works often have disregarded information at occluded regions. However, we argue that occluded regions have strong correlations with hands so that they can provide highly beneficial information for complete 3D hand mesh estimation. Thus, in this work, we propose a novel 3D hand mesh estimation network HandOccNet, that can fully exploits the information at occluded regions as a secondary means to enhance image features and make it much richer. To this end, we design two successive Transformer-based modules, called feature injecting

transformer (FIT) and self-enhancing transformer (SET). FIT injects hand information into occluded region by considering their correlation. SET refines the output of FIT by using a self-attention mechanism. By injecting the hand information to the occluded region, our HandOccNet reaches the state-of-the-art performance on 3D hand mesh benchmarks that contain challenging hand-object occlusions. The codes are available in: https://github.com/namepllet/HandOccNet.
****

EfficientNeRF Efficient Neural Radiance Fields
Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12902-12911
Neural Radiance Fields (NeRF) has been wildly applied to various tasks for its high-quality representation of 3D scenes. It takes long per-scene training time and per-image testing time. In this paper, we present EfficientNeRF as an efficient NeRF-based method to represent 3D scene and synthesize novel-view images. Although several ways exist to accelerate the training or testing process, it is still difficult to much reduce time for both phases simultaneously. We analyze the density and weight distribution of the sampled points then propose valid and pivotal sampling at the coarse and fine stage, respectively, to significantly improve sampling efficiency. In addition, we design a novel data structure to cache the whole scene during testing to accelerate the testing speed. Overall, our method can reduce over 88% of training time, reach testing speed of around 200 to 500 FPS, while still achieving competitive accuracy. Experiments prove that our method promotes the practicality of NeRF in the real world and enables many applications.
****

Quantifying Societal Bias Amplification in Image Captioning
Yusuke Hirota, Yuta Nakashima, Noa Garcia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13450-13459
We study societal bias amplification in image captioning. Image captioning models have been shown to perpetuate gender and racial biases, however, metrics to measure, quantify, and evaluate the societal bias in captions are not yet standardized. We provide a comprehensive study on the strengths and limitations of each metric, and propose LIC, a metric to study captioning bias amplification. We argue that, for image captioning, it is not enough to focus on the correct prediction of the protected attribute, and the whole context should be taken into account. We conduct extensive evaluation on traditional and state-of-the-art image captioning models, and surprisingly find that, by only focusing on the protected attribute prediction, bias mitigation models are unexpectedly amplifying bias.
****

Modular Action Concept Grounding in Semantic Video Prediction
Wei Yu, Wenxin Chen, Songheng Yin, Steve Easterbrook, Animesh Garg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3605-3614
Recent works in video prediction have mainly focused on passive forecasting and low-level action-conditional prediction, which sidesteps the learning of interaction between agents and objects. We introduce the task of semantic action-conditional video prediction, which uses semantic action labels to describe those interactions and can be regarded as an inverse problem of action recognition. The challenge of this new task primarily lies in how to effectively inform the model of semantic action information. Inspired by the idea of Mixture of Experts, we embody each abstract label by a structured combination of various visual concept learners and propose a novel video prediction model, Modular Action Concept Network (MAC). Our method is evaluated on two newly designed synthetic datasets, CLEVR-Building-Blocks and Sapien-Kitchen, and one real-world dataset called Tower-Creation. Extensive experiments demonstrate that MAC can correctly condition on given instructions and generate corresponding future frames without need of bounding boxes. We further show that the trained model can make out-of-distribution generalization, be quickly adapted to new object categories and exploit its learnt features for object detection, showing the progression towards higher-level cog

nitive abilities. More visualizations can be found at http://www.pair.toronto.edu/mac/.
****

StyleSwin: Transformer-Based GAN for High-Resolution Image Generation
Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11304-11314
Despite the tantalizing success in a broad of vision tasks, transformers have not yet demonstrated on-par ability as ConvNets in high-resolution image generative modeling. In this paper, we seek to explore using pure transformers to build a generative adversarial network for high-resolution image synthesis. To this end, we believe that the local attention is crucial to strike the balance between computational efficiency and modeling capacity. Hence, the proposed generator adopts Swin transformer in a style-based architecture. To achieve larger receptive field, we propose double attention which simultaneously leverages the context of the local and the shifted windows, leading to improved generation quality. Moreover, we show that offering the knowledge of the absolute position that has lost in window-based transformers greatly benefits the generation quality. The proposed StyleSwin is scalable to high resolutions, with both the coarse geometry and fine structures benefit from the strong expressivity of transformers. However, blocking artifacts occur during high-resolution synthesis because performing the local attention in a block-wise manner may break the spatial coherency. To solve this, we empirically investigate various solutions, among which we find that employing a wavelet discriminator to examine the spectral discrepancy effectively suppresses the artifacts. Extensive experiments show the superiority over prior transformer-based GANs, especially on high resolutions, e.g., 1024x1024. The StyleSwin, without complex training strategies, excelling over StyleGAN on CelebA-HQ 1024, and achieves on-par performance on FFHQ-1024, proving the promise of using transformers for high-resolution image generation. The code and pretrained models are available at https://github.com/microsoft/StyleSwin.
****

Reinforced Structured State-Evolution for Vision-Language Navigation
Jinyu Chen, Chen Gao, Erli Meng, Qiong Zhang, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15450-15459
Vision-and-language Navigation (VLN) task requires an embodied agent to navigate to a remote location following a natural language instruction. Previous methods usually adopt a sequence model (e.g., Transformer and LSTM) as the navigator. In such a paradigm, the sequence model predicts action at each step through a maintained navigation state, which is generally represented as a one-dimensional vector. However, the crucial navigation clues (i.e., object-level environment layout) for embodied navigation task is discarded since the maintained vector is essentially unstructured. In this paper, we propose a novel Structured state-Evolution (SEvol) model to effectively maintain the environment layout clues for VLN. Specifically, we utilise the graph-based feature to represent the navigation state instead of the vector-based state. Accordingly, we devise a Reinforced Layout clues Miner (RLM) to mine and detect the most crucial layout graph for long-term navigation via a customised reinforcement learning strategy. Moreover, the Structured Evolving Module (SEM) is proposed to maintain the structured graph-based state during navigation, where the state is gradually evolved to learn the object-level spatial-temporal relationship. The experiments on the R2R and R4R datasets show that the proposed SEvol model improves VLN models' performance by large margins, e.g., +3% absolute SPL accuracy for NvEM and +8% for EnvDrop on the R2R test set.
****

Sub-Word Level Lip Reading With Visual Attention
K R Prajwal, Triantafyllos Afouras, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5162-5172
The goal of this paper is to learn strong lip reading models that can recognise

speech in silent videos. Most prior works deal with the open-set visual speech recognition problem by adapting existing automatic speech recognition techniques on top of trivially pooled visual features. Instead, in this paper, we focus on the unique challenges encountered in lip reading and propose tailored solutions. To this end, we make the following contributions: (1) we propose an attention-based pooling mechanism to aggregate visual speech representations; (2) we use sub-word units for lip reading for the first time and show that this allows us to better model the ambiguities of the task; (3) we propose a model for Visual Speech Detection (VSD), trained on top of the lip reading network. Following the above, we obtain state-of-the-art results on the challenging LRS2 and LRS3 benchmarks when training on public datasets, and even surpass models trained on large-scale industrial datasets by using an order of magnitude less data. Our best model achieves 22.6% word error rate on the LRS2 dataset, a performance unprecedented for lip reading models, significantly reducing the performance gap between lip reading and automatic speech recognition. Moreover, on the AVA-ActiveSpeaker benchmark, our VSD model surpasses all visual-only baselines and even outperforms several recent audio-visual methods.

****

Weakly Supervised High-Fidelity Clothing Model Generation

Ruili Feng, Cheng Ma, Chengji Shen, Xin Gao, Zhenjiang Liu, Xiaobo Li, Kairi Ou, Deli Zhao, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3440-3449

The development of online economics arouses the demand of generating images of models on product clothes, to display new clothes and promote sales. However, the expensive proprietary model images challenge the existing image virtual try-on methods in this scenario, as most of them need to be trained on considerable amounts of model images accompanied with paired clothes images. In this paper, we propose a cheap yet scalable weakly-supervised method called Deep Generative Projection (DGP) to address this specific scenario. Lying in the heart of the proposed method is to imitate the process of human predicting the wearing effect, which is an unsupervised imagination based on life experience rather than computation rules learned from supervisions. Here a pretrained StyleGAN is used to capture the practical experience of wearing. Experiments show that projecting the rough alignment of clothing and body onto the StyleGAN space can yield photo-realistic wearing results. Experiments on real scene proprietary model images demonstrate the superiority of DGP over several state-of-the-art supervised methods when generating clothing model images.

****

Highly-Efficient Incomplete Large-Scale Multi-View Clustering With Consensus Bipartite Graph

Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, En Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9776-9785

Multi-view clustering has received increasing attention due to its effectiveness in fusing complementary information without manual annotations. Most previous methods hold the assumption that each instance appears in all views. However, it is not uncommon to see that some views may contain some missing instances, which gives rise to incomplete multi-view clustering (IMVC) in literature. Although many IMVC methods have been recently proposed, they always encounter high complexity and expensive time expenditure from being applied into large-scale tasks. In this paper, we present a flexible highly-efficient incomplete large-scale multi-view clustering approach based on bipartite graph framework to solve these issues. Specifically, we formalize multi-view anchor learning and incomplete bipartite graph into a unified framework, which coordinates with each other to boost cluster performance. By introducing the flexible bipartite graph framework to handle IMVC for the first practice, our proposed method enjoys linear complexity respecting to instance numbers, which is more applicable for large-scale IMVC tasks. Comprehensive experimental results on various benchmark datasets demonstrate the effectiveness and efficiency of our proposed algorithm against other IMVC competitors.

****
Towards Principled Disentanglement for Domain Generalization
Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8024-8034
A fundamental challenge for machine learning models is generalizing to out-of-distribution (OOD) data, in part due to spurious correlations. To tackle this challenge, we first formalize the OOD generalization problem as constrained optimization, called Disentanglement-constrained Domain Generalization (DDG). We relax this non-trivial constrained optimization problem to a tractable form with finite-dimensional parameterization and empirical approximation. Then a theoretical analysis of the extent to which the above transformations deviates from the original problem is provided. Based on the transformation, we propose a primal-dual algorithm for joint representation disentanglement and domain generalization. In contrast to traditional approaches based on domain adversarial training and domain labels, DDG jointly learns semantic and variation encoders for disentanglement, enabling flexible manipulation and augmentation on training data. DDG aims to learn intrinsic representations of semantic concepts that are invariant to nuisance factors and generalizable across domains. Comprehensive experiments on popular benchmarks show that DDG can achieve competitive OOD performance and uncover interpretable salient structures within data.
****
Discrete Cosine Transform Network for Guided Depth Map Super-Resolution
Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, Hanspeter Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5697-5707
Guided depth super-resolution (GDSR) is an essential topic in multi-modal image processing, which reconstructs high-resolution (HR) depth maps from low-resolution ones collected with suboptimal conditions with the help of HR RGB images of the same scene. To solve the challenges in interpreting the working mechanism, extracting cross-modal features and RGB texture over-transferred, we propose a novel Discrete Cosine Transform Network (DCTNet) to alleviate the problems from three aspects. First, the Discrete Cosine Transform (DCT) module reconstructs the multi-channel HR depth features by using DCT to solve the channel-wise optimization problem derived from the image domain. Second, we introduce a semi-coupled feature extraction module that uses shared convolutional kernels to extract common information and private kernels to extract modality-specific information. Third, we employ an edge attention mechanism to highlight the contours informative for guided upsampling. Extensive quantitative and qualitative evaluations demonstrate the effectiveness of our DCTNet, which outperforms previous state-of-the-art methods with a relatively small number of parameters. Codes are available at https://github.com/Zhaozixiang1228/GDSR-DCTNet.
****
Cerberus Transformer: Joint Semantic, Affordance and Attribute Parsing
Xiaoxue Chen, Tianyu Liu, Hao Zhao, Guyue Zhou, Ya-Qin Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19649-19658
Multi-task indoor scene understanding is widely considered as an intriguing formulation, as the affinity of different tasks may lead to improved performance. In this paper, we tackle the new problem of joint semantic, affordance and attribute parsing. However, successfully resolving it requires a model to capture long-range dependency, learn from weakly aligned data and properly balance sub-tasks during training. To this end, we propose an attention-based architecture named Cerberus and a tailored training framework. Our method effectively addresses aforementioned challenges and achieves state-of-the-art performance on all three tasks. Moreover, an in-depth analysis shows concept affinity consistent with human cognition, which inspires us to explore the possibility of extremely low-shot learning. Surprisingly, Cerberus achieves strong results using only 0.1%-1% annotation. Visualizations further confirm that this success is credited to common attention maps across tasks. Code and models can be accessed at https://github.com/

OPEN-AIR-SUN/Cerberus.
****
****
CoSSL: Co-Learning of Representation and Classifier for Imbalanced Semi-Supervised Learning
Yue Fan, Dengxin Dai, Anna Kukleva, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14574-14584
Standard semi-supervised learning (SSL) using class-balanced datasets has shown great progress to leverage unlabeled data effectively. However, the more realistic setting of class-imbalanced data - called imbalanced SSL - is largely underexplored and standard SSL tends to underperform. In this paper, we propose a novel co-learning framework (CoSSL), which decouples representation and classifier learning while coupling them closely. To handle the data imbalance, we devise Tail-class Feature Enhancement (TFE) for classifier learning. Furthermore, the current evaluation protocol for imbalanced SSL focuses only on balanced test sets, which has limited practicality in real-world scenarios. Therefore, we further conduct a comprehensive evaluation under various shifted test distributions. In experiments, we show that our approach outperforms other methods over a large range of shifted distributions, achieving state-of-the-art performance on benchmark datasets ranging from CIFAR-10, CIFAR-100, ImageNet, to Food-101. Our code will be made publicly available.
****
Discovering Objects That Can Move
Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, Martial Hebert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11789-11798
This paper studies the problem of object discovery -- separating objects from the background without manual labels. Existing approaches utilize appearance cues, such as color, texture, and location, to group pixels into object-like regions. However, by relying on appearance alone, these methods fail to separate objects from the background in cluttered scenes. This is a fundamental limitation since the definition of an object is inherently ambiguous and context-dependent. To resolve this ambiguity, we choose to focus on dynamic objects -- entities that can move independently in the world. We then scale the recent auto-encoder based frameworks for unsupervised object discovery from toy synthetic images to complex real-world scenes. To this end, we simplify their architecture, and augment the resulting model with a weak learning signal from general motion segmentation algorithms. Our experiments demonstrate that, despite only capturing a small subset of the objects that move, this signal is enough to generalize to segment both moving and static instances of dynamic objects. We show that our model scales to a newly collected, photo-realistic synthetic dataset with street driving scenarios. Additionally, we leverage ground truth segmentation and flow annotations in this dataset for thorough ablation and evaluation. Finally, our experiments on the real-world KITTI benchmark demonstrate that the proposed approach outperforms both heuristic- and learning-based methods by capitalizing on motion cues.
****
Knowledge Mining With Scene Text for Fine-Grained Recognition
Hao Wang, Junchao Liao, Tianheng Cheng, Zewen Gao, Hao Liu, Bo Ren, Xiang Bai, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4624-4633
Recently, the semantics of scene text has been proven to be essential in fine-grained image classification. However, the existing methods mainly exploit the literal meaning of scene text for fine-grained recognition, which might be irrelevant when it is not significantly related to objects/scenes. We propose an end-to-end trainable network that mines implicit contextual knowledge behind scene text image and enhance the semantics and correlation to fine-tune the image representation. Unlike the existing methods, our model integrates three modalities: visual feature extraction, text semantics extraction, and correlating background knowledge to fine-grained image classification. Specifically, we employ KnowBert to

retrieve relevant knowledge for semantic representation and combine it with ima
ge features for fine-grained classification. Experiments on two benchmark datase
ts, Con-Text, and Drink Bottle, show that our method outperforms the state-of-th
e-art by 3.72% mAP and 5.39% mAP, respectively. To further validate the effectiv
eness of the proposed method, we create a new dataset on crowd activity recognit
ion for the evaluation. The source code, new dataset, and pre-trained models of
this work will be publicly available.
****

Self-Supervised Learning of Object Parts for Semantic Segmentation
Adrian Ziegler, Yuki M. Asano; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2022, pp. 14502-14511
Progress in self-supervised learning has brought strong general image representa
tion learning methods. Yet so far, it has mostly focused on image-level learning
. In turn, tasks such as unsupervised image segmentation have not benefited from
this trend as they require spatially-diverse representations. However, learning
dense representations is challenging, as in the unsupervised context it is not
clear how to guide the model to learn representations that correspond to various
potential object categories. In this paper, we argue that self-supervised learn
ing of object parts is a solution to this issue. Object parts are generalizable:
they are a priori independent of an object definition, but can be grouped to fo
rm objects a posteriori. To this end, we leverage the recently proposed Vision T
ransformer's capability of attending to objects and combine it with a spatially
dense clustering task for fine-tuning the spatial tokens. Our method surpasses t
he state-of-the-art on three semantic segmentation benchmarks by 17%-3%, showing
that our representations are versatile under various object definitions. Finall
y, we extend this to fully unsupervised segmentation - which refrains completely
from using label information even at test-time - and demonstrate that a simple
method for automatically merging discovered object parts based on community dete
ction yields substantial gains.
****

Iterative Corresponding Geometry: Fusing Region and Depth for Highly Efficient 3
D Tracking of Textureless Objects
Manuel Stoiber, Martin Sundermeyer, Rudolph Triebel; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6855-68
65
Tracking objects in 3D space and predicting their 6DoF pose is an essential task
in computer vision. State-of-the-art approaches often rely on object texture to
tackle this problem. However, while they achieve impressive results, many objec
ts do not contain sufficient texture, violating the main underlying assumption.
In the following, we thus propose ICG, a novel probabilistic tracker that fuses
region and depth information and only requires the object geometry. Our method d
eploys correspondence lines and points to iteratively refine the pose. We also i
mplement robust occlusion handling to improve performance in real-world settings
. Experiments on the YCB-Video, OPT, and Choi datasets demonstrate that, even fo
r textured objects, our approach outperforms the current state of the art with r
espect to accuracy and robustness. At the same time, ICG shows fast convergence
and outstanding efficiency, requiring only 1.3 ms per frame on a single CPU core
. Finally, we analyze the influence of individual components and discuss our per
formance compared to deep learning-based methods. The source code of our tracker
is publicly available.
****

Single-Photon Structured Light
Varun Sundar, Sizhuo Ma, Aswin C. Sankaranarayanan, Mohit Gupta; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
pp. 17865-17875
We present a novel structured light technique that uses Single Photon Avalanche
Diode (SPAD) arrays to enable 3D scanning at high-frame rates and low-light leve
ls. This technique, called "Single-Photon Structured Light", works by sensing bi
nary images that indicates the presence or absence of photon arrivals during eac
h exposure; the SPAD array is used in conjunction with a high-speed binary proje

ctor, with both devices operated at speeds as high as 20 kHz. The binary images that we acquire are heavily influenced by photon noise and are easily corrupted by ambient sources of light. To address this, we develop novel temporal sequences using error correction codes that are designed to be robust to short-range effects like projector and camera defocus as well as resolution mismatch between the two devices. Our lab prototype is capable of 3D imaging in challenging scenarios involving objects with extremely low albedo or undergoing fast motion, as well as scenes under strong ambient illumination.
****

Deblurring via Stochastic Refinement
Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G. Dimakis, Peyman Milanfar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16293-16303
Image deblurring is an ill-posed problem with multiple plausible solutions for a given input image. However, most existing methods produce a deterministic estimate of the clean image and are trained to minimize pixel-level distortion. These metrics are known to be poorly correlated with human perception, and often lead to unrealistic reconstructions. We present an alternative framework for blind deblurring based on conditional diffusion models. Unlike existing techniques, we train a stochastic sampler that refines the output of a deterministic predictor and is capable of producing a diverse set of plausible reconstructions for a given input. This leads to a significant improvement in perceptual quality over existing state-of-the-art methods across multiple standard benchmarks. Our predict-and-refine approach also enables much more efficient sampling compared to typical diffusion models. Combined with a carefully tuned network architecture and inference procedure, our method is competitive in terms of distortion metrics such as PSNR. These results show clear benefits of our diffusion-based method for deblurring and challenge the widely used strategy of producing a single, deterministic reconstruction.
****

3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds
Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16464-16473
Observing that the 3D captioning task and the 3D grounding task contain both shared and complementary information in nature, in this work, we propose a unified framework to jointly solve these two distinct but closely related tasks in a synergistic fashion, which consists of both shared task-agnostic modules and lightweight task-specific modules. On one hand, the shared task-agnostic modules aim to learn precise locations of objects, fine-grained attribute features to characterize different objects, and complex relations between objects, which benefit both captioning and visual grounding. On the other hand, by casting each of the two tasks as the proxy task of another one, the lightweight task-specific modules solve the captioning task and the grounding task respectively. Extensive experiments and ablation study on three 3D vision and language datasets demonstrate that our joint training framework achieves significant performance gains for each individual task and finally improves the state-of-the-art performance for both captioning and grounding tasks.
****

TransGeo: Transformer Is All You Need for Cross-View Image Geo-Localization
Sijie Zhu, Mubarak Shah, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1162-1171
The dominant CNN-based methods for cross-view image geo-localization rely on polar transform and fail to model global correlation. We propose a pure transformer-based approach (TransGeo) to address these limitations from a different perspective. TransGeo takes full advantage of the strengths of transformer related to global information modeling and explicit position information encoding. We further leverage the flexibility of transformer input and propose an attention-guided non-uniform cropping method, so that uninformative image patches are removed wit

h negligible drop on performance to reduce computation cost. The saved computati on can be reallocated to increase resolution only for informative patches, resul ting in performance improvement with no additional computation cost. This "atten d and zoom-in" strategy is highly similar to human behavior when observing image s. Remarkably, TransGeo achieves state-of-the-art results on both urban and rura l datasets, with significantly less computation cost than CNN-based methods. It does not rely on polar transform and infers faster than CNN-based methods. Code is available at https://github.com/Jeff-Zilence/TransGeo2022.
****

R(Det)2: Randomized Decision Routing for Object Detection
Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 4825-4834
In the paradigm of object detection, the decision head is an important part, whi ch affects detection performance significantly. Yet how to design a high-perform ance decision head remains to be an open issue. In this paper, we propose a nove l approach to combine decision trees and deep neural networks in an end-to-end l earning manner for object detection. First, we disentangle the decision choices and prediction values by plugging soft decision trees into neural networks. To f acilitate the effective learning, we propose the randomized decision routing wit h node selective and associative losses, which can boost the feature representat ive learning and network decision simultaneously. Second, we develop the decisio n head for object detection with narrow branches to generate the routing probabi lities and masks, for the purpose of obtaining divergent decisions from differen t nodes. We name this approach as the randomized decision routing for object det ection, abbreviated as R(Det)^2. Experiments on MS-COCO dataset demonstrate that R(Det)^2 is effective to improve the detection performance. Equipped with exist ing detectors, it achieves 1.4~ 3.6% AP improvement. Code will be released soon.
****

Abandoning the Bayer-Filter To See in the Dark
Xingbo Dong, Wanyan Xu, Zhihui Miao, Lan Ma, Chao Zhang, Jiewen Yang, Zhe Jin, A ndrew Beng Jin Teoh, Jiajun Shen; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2022, pp. 17431-17440
Low-light image enhancement, a pervasive but challenging problem, plays a centra l role in enhancing the visibility of an image captured in a poor illumination e nvironment. Due to the fact that not all photons can pass the Bayer-Filter on th e sensor of the color camera, in this work, we first present a De-Bayer-Filter s imulator based on deep neural networks to generate a monochrome raw image from t he colored raw image. Next, a fully convolutional network is proposed to achieve the low-light image enhancement by fusing colored raw data with synthesized mon ochrome data. Channel-wise attention is also introduced to the fusion process to establish a complementary interaction between features from colored and monochr ome raw images. To train the convolutional networks, we propose a dataset with m onochrome and color raw pairs named Mono-Colored Raw paired dataset (MCR) collec ted by using a monochrome camera without Bayer-Filter and a color camera with Ba yer-Filter. The proposed pipeline take advantages of the fusion of the virtual m onochrome and the color raw images and our extensive experiments indicate that s ignificant improvement can be achieved by leveraging raw sensor data and data-dr iven learning.
****

SASIC: Stereo Image Compression With Latent Shifts and Stereo Attention
Matthias Wödlinger, Jan Kotera, Jan Xu, Robert Sablatnig; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 66 1-670
We propose a learned method for stereo image compression that leverages the simi larity of the left and right images in a stereo pair due to overlapping fields o f view. The left image is compressed by a learned compression method based on an autoencoder with a hyperprior entropy model. The right image uses this informat ion from the previously encoded left image in both the encoding and decoding sta ges. In particular, for the right image, we encode only the residual of its late nt representation to the optimally shifted latent of the left image. On top of t

hat, we also employ a stereo attention module to connect left and right images d
uring decoding. The performance of the proposed method is evaluated on two bench
mark stereo image datasets (Cityscapes and InStereo2K) and outperforms previous
stereo image compression methods while being significantly smaller in model size
.
****

Exploiting Temporal Relations on Radar Perception for Autonomous Driving
Peizhao Li, Pu Wang, Karl Berntorp, Hongfu Liu; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17071-17080
We consider the object recognition problem in autonomous driving using automotiv
e radar sensors. Comparing to Lidar sensors, radar is cost-effective and robust
in all-weather conditions for perception in autonomous driving. However, radar s
ignals suffer from low angular resolution and precision in recognizing surroundi
ng objects. To enhance the capacity of automotive radar, in this work, we exploi
t the temporal information from successive ego-centric bird-eye-view radar image
 frames for radar object recognition. We leverage the consistency of an object's
 existence and attributes (size, orientation, etc.), and propose a temporal rela
tional layer to explicitly model the relations between objects within successive
 radar images. In both object detection and multiple object tracking, we show th
e superiority of our method compared to several baseline approaches.
****

Multi-Instance Point Cloud Registration by Efficient Correspondence Clustering
Weixuan Tang, Danping Zou; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2022, pp. 6667-6676
We address the problem of estimating the poses of multiple instances of the sour
ce point cloud within a target point cloud. Existing solutions require sampling
a lot of hypotheses to detect possible instances and reject the outliers, whose
robustness and efficiency degrade notably when the number of instances and outli
ers increase. We propose to directly group the set of noisy correspondences into
 different clusters based on a distance invariance matrix. The instances and out
liers are automatically identified through clustering. Our method is robust and
fast. We evaluated our method on both synthetic and real-world datasets. The res
ults show that our approach can correctly register up to 20 instances with an F1
 score of 90.46% in the presence of 70% outliers, which performs significantly b
etter and at least 10x faster than existing methods.
****

Contrastive Boundary Learning for Point Cloud Segmentation
Liyao Tang, Yibing Zhan, Zhe Chen, Baosheng Yu, Dacheng Tao; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp.
 8489-8499
Point cloud segmentation is fundamental in understanding 3D environments. Howeve
r, current 3D point cloud segmentation methods usually perform poorly on scene b
oundaries, which degenerates the overall segmentation performance. In this paper
, we focus on the segmentation of scene boundaries. Accordingly, we first explor
e metrics to evaluate the segmentation performance on scene boundaries. To addre
ss the unsatisfactory performance on boundaries, we then propose a novel contras
tive boundary learning (CBL) framework for point cloud segmentation. Specificall
y, the proposed CBL enhances feature discrimination between points across bounda
ries by contrasting their representations with the assistance of scene contexts
at multiple scales. By applying CBL on three different baseline methods, we expe
rimentally show that CBL consistently improves different baselines and assists t
hem to achieve compelling performance on boundaries, as well as the overall perf
ormance, e.g. in mIoU. The experimental results demonstrate the effectiveness of
 our method and the importance of boundaries for 3D point cloud segmentation. Co
de and model will be made publicly available at https://github.com/LiyaoTang/con
trastBoundary.
****

Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Im
age Super-Resolution
Jie Liang, Hui Zeng, Lei Zhang; Proceedings of the IEEE/CVF Conference on Comput

er Vision and Pattern Recognition (CVPR), 2022, pp. 5657-5666
Single image super-resolution (SISR) with generative adversarial networks (GAN) has recently attracted increasing attention due to its potentials to generate rich details. However, the training of GAN is unstable, and it often introduces many perceptually unpleasant artifacts along with the generated details. In this paper, we demonstrate that it is possible to train a GAN-based SISR model which can stably generate perceptually realistic details while inhibiting visual artifacts. Based on the observation that the local statistics (e.g., residual variance) of artifact areas are often different from the areas of perceptually friendly details, we develop a framework to discriminate between GAN-generated artifacts and realistic details, and consequently generate an artifact map to regularize and stabilize the model training process. Our proposed locally discriminative learning (LDL) method is simple yet effective, which can be easily plugged in off-the-shelf SISR methods and boost their performance. Experiments demonstrate that LDL outperforms the state-of-the-art GAN based SISR methods, achieving not only higher reconstruction accuracy but also superior perceptual quality on both synthetic and real-world datasets. Codes and models are available at https://github.com/csjliang/LDL.
****

CVNet: Contour Vibration Network for Building Extraction
Ziqiang Xu, Chunyan Xu, Zhen Cui, Xiangwei Zheng, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1383-1391
The classic active contour model raises a great promising solution to polygon-based object extraction with the progress of deep learning recently. Inspired by the physical vibration theory, we propose a contour vibration network (CVNet) for automatic building boundary delineation. Different from the previous contour models, the CVNet originally roots in the force and motion principle of contour string. Through the infinitesimal analysis and Newton's second law, we derive the spatial-temporal contour vibration model of object shapes, which is mathematically reduced to second-order differential equation. To concretize the dynamic model, we transform the vibration model into the space of image features, and reparameterize the equation coefficients as the learnable state from feature domain. The contour changes are finally evolved in a progressive mode through the computation of contour vibration equation. Both the polygon contour evolution and the model optimization are modulated to form a close-looping end-to-end network. Comprehensive experiments on three datasets demonstrate the effectiveness and superiority of our CVNet over other baselines and state-of-the-art methods for the polygon-based building extraction. The code is available at https://github.com/xzq-njust/CVNet.
****

Hyperbolic Image Segmentation
Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, Pascal Mettes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4453-4462
For image segmentation, the current standard is to perform pixel-level optimization and inference in Euclidean output embedding spaces through linear hyperplanes. In this work, we show that hyperbolic manifolds provide a valuable alternative for image segmentation and propose a tractable formulation of hierarchical pixel-level classification in hyperbolic space. Hyperbolic Image Segmentation opens up new possibilities and practical benefits for segmentation, such as uncertainty estimation and boundary information for free, zero-label generalization, and increased performance in low-dimensional output embeddings.
****

Forward Compatible Training for Large-Scale Embedding Retrieval Systems
Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, Hadi Pouransari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19386-19395
In visual retrieval systems, updating the embedding model requires recomputing features for every piece of data. This expensive process is referred to as backfi

lling. Recently, the idea of backward compatible training (BCT) was proposed. To avoid the cost of backfilling, BCT modifies training of the new model to make its representations compatible with those of the old model. However, BCT can significantly hinder the performance of the new model. In this work, we propose a new learning paradigm for representation learning: forward compatible training (FCT). In FCT, when the old model is trained, we also prepare for a future unknown version of the model. We propose learning side-information, an auxiliary feature for each sample which facilitates future updates of the model. To develop a powerful and flexible framework for model compatibility, we combine side-information with a forward transformation from old to new embeddings. Training of the new model is not modified, hence, its accuracy is not degraded. We demonstrate significant retrieval accuracy improvement compared to BCT for various datasets: ImageNet-1k (+18.1%), Places-365 (+5.4%), and VGG-Face2 (+8.3%). FCT obtains model compatibility when the new and old models are trained across different datasets, losses, and architectures.

****

Everything at Once - Multi-Modal Fusion Transformer for Video Retrieval
Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S. Feris, David Harwath, James Glass, Hilde Kuehne; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20020-20029

Multi-modal learning from video data has seen increased attention recently as it allows training of semantically meaningful embeddings without human annotation, enabling tasks like zero-shot retrieval and action localization. In this work, we present a multi-modal, modality agnostic fusion transformer that learns to exchange information between multiple modalities, such as video, audio, and text, and integrate them into a fused representation in a joined multi-modal embedding space. We propose to train the system with a combinatorial loss on everything at once - any combination of input modalities, such as single modalities as well as pairs of modalities, explicitly leaving out any add-ons such as position or modality encoding. At test time, the resulting model can process and fuse any number of input modalities. Moreover, the implicit properties of the transformer allow to process inputs of different lengths. To evaluate the proposed approach, we train the model on the large scale HowTo100M dataset and evaluate the resulting embedding space on four challenging benchmark datasets obtaining state-of-the-art results in zero-shot video retrieval and zero-shot video action localization. Our code for this work is also available.

****

Swin Transformer V2: Scaling Up Capacity and Resolution
Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12009-12019

We present techniques for scaling Swin Transformer [??] up to 3 billion parameters and making it capable of training with images of up to 1,536x1,536 resolution. By scaling up capacity and resolution, Swin Transformer sets new records on four representative vision benchmarks: 84.0% top-1 accuracy on ImageNet-V2 image classification, 63.1 / 54.4 box / mask mAP on COCO object detection, 59.9 mIoU on ADE20K semantic segmentation, and 86.8% top-1 accuracy on Kinetics-400 video action classification. We tackle issues of training instability, and study how to effectively transfer models pre-trained at low resolutions to higher resolution ones. To this aim, several novel technologies are proposed: 1) a residual post normalization technique and a scaled cosine attention approach to improve the stability of large vision models; 2) a log-spaced continuous position bias technique to effectively transfer models pre-trained at low-resolution images and windows to their higher-resolution counterparts. In addition, we share our crucial implementation details that lead to significant savings of GPU memory consumption and thus make it feasible to train large vision models with regular GPUs. Using these techniques and self-supervised pre-training, we successfully train a strong 3 billion Swin Transformer model and effectively transfer it to various vision

tasks involving high-resolution images or windows, achieving the state-of-the-art accuracy on a variety of benchmarks. Code is available at https://github.com/microsoft/Swin-Transformer.
****

Neural Template: Topology-Aware Reconstruction and Disentangled Generation of 3D Meshes
Ka-Hei Hui, Ruihui Li, Jingyu Hu, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18572-18582
This paper introduces a novel framework called DT-Net for 3D mesh reconstruction and generation via Disentangled Topology. Beyond previous works, we learn a topology-aware neural template specific to each input then deform the template to reconstruct a detailed mesh while preserving the learned topology. One key insight is to decouple the complex mesh reconstruction into two sub-tasks: topology formulation and shape deformation. Thanks to the decoupling, DT-Net implicitly learns a disentangled representation for the topology and shape in the latent space. Hence, it can enable novel disentangled controls for supporting various shape generation applications, eg, remix the topologies of 3D objects, that are not achievable by previous reconstruction works. Extensive experimental results demonstrate that our method is able to produce high-quality meshes, particularly with diverse topologies, as compared with the state-of-the-art methods.
****

DEFEAT: Deep Hidden Feature Backdoor Attacks by Imperceptible Perturbation and Latent Representation Constraints
Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, Kaitai Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15213-15222
Backdoor attack is a type of serious security threat to deep learning models.An adversary can provide users with a model trained on poisoned data to manipulate prediction behavior in test stage using a backdoor. The backdoored models behave normally on clean images, yet can be activated and output incorrect prediction if the input is stamped with a specific trigger pattern.Most existing backdoor attacks focus on manually defining imperceptible triggers in input space without considering the abnormality of triggers' latent representations in the poisoned model.These attacks are susceptible to backdoor detection algorithms and even visual inspection.In this paper, We propose a novel and stealthy backdoor attack - DEFEAT. It poisons the clean data using adaptive imperceptible perturbation and restricts latent representation during training process to strengthen our attack's stealthiness and resistance to defense algorithms.We conduct extensive experiments on multiple image classifiers using real-world datasets to demonstrate that our attack can 1) hold against the state-of-the-art defenses, 2) deceive the victim model with high attack success without jeopardizing model utility, and 3) provide practical stealthiness on image data.
****

Projective Manifold Gradient Layer for Deep Rotation Regression
Jiayi Chen, Yingda Yin, Tolga Birdal, Baoquan Chen, Leonidas J. Guibas, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6646-6655
Regressing rotations on SO(3) manifold using deep neural networks is an important yet unsolved problem. The gap between the Euclidean network output space and the non-Euclidean SO(3) manifold imposes a severe challenge for neural network learning in both forward and backward passes. While several works have proposed different regression-friendly rotation representations, very few works have been devoted to improving the gradient backpropagating in the backward pass. In this paper, we propose a manifold-aware gradient that directly backpropagates into deep network weights. Leveraging Riemannian optimization to construct a novel projective gradient, our proposed regularized projective manifold gradient (RPMG) method helps networks achieve new state-of-the-art performance in a variety of rotation estimation tasks. Our proposed gradient layer can also be applied to other smooth manifolds such as the unit sphere. Our project page is at https://jychen18.github.io/RPMG.

****
CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation
Jinheng Xie, Xianxu Hou, Kai Ye, Linlin Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4483-4492
It has been widely known that CAM (Class Activation Map) usually only activates discriminative object regions and falsely includes lots of object-related backgrounds. As only a fixed set of image-level object labels are available to the WSSS (weakly supervised semantic segmentation) model, it could be very difficult to suppress those diverse background regions consisting of open set objects. In this paper, we propose a novel Cross Language Image Matching (CLIMS) framework, based on the recently introduced Contrastive Language-Image Pre-training (CLIP) model, for WSSS. The core idea of our framework is to introduce natural language supervision to activate more complete object regions and suppress closely-related open background regions. In particular, we design object, background region and text label matching losses to guide the model to excite more reasonable object regions for CAM of each category. In addition, we design a co-occurring background suppression loss to prevent the model from activating closely-related background regions, with a predefined set of class-related background text descriptions. These designs enable the proposed CLIMS to generate a more complete and compact activation map for the target objects. Extensive experiments on PASCAL VOC2012 dataset show that our CLIMS significantly outperforms the previous state-of-the-art methods.
****
Learning To Refactor Action and Co-Occurrence Features for Temporal Action Localization
Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, Wei Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13884-13893
The main challenge of Temporal Action Localization is to retrieve subtle human actions from various co-occurring ingredients, e.g., context and background, in an untrimmed video. While prior approaches have achieved substantial progress through devising advanced action detectors, they still suffer from these co-occurring ingredients which often dominate the actual action content in videos. In this paper, we explore two orthogonal but complementary aspects of a video snippet, i.e., the action features and the co-occurrence features. Especially, we develop a novel auxiliary task by decoupling these two types of features within a video snippet and recombining them to generate a new feature representation with more salient action information for accurate action localization. We term our method RefactorNet, which first explicitly factorizes the action content and regularizes its co-occurrence features, and then synthesizes a new action-dominated video representation. Extensive experimental results and ablation studies on THUMOS14 and ActivityNet v1.3 demonstrate that our new representation, combined with a simple action detector, can significantly improve the action localization performance.
****
It's Time for Artistic Correspondence in Music and Video
Dídac Surís, Carl Vondrick, Bryan Russell, Justin Salamon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10564-10574
We present an approach for recommending a music track for a given video, and vice versa, based on both their temporal alignment and their correspondence at an artistic level. We propose a self-supervised approach that learns this correspondence directly from data, without any need of human annotations. In order to capture the high-level concepts that are required to solve the task, we propose modeling the long-term temporal context of both the video and the music signals, using Transformer networks for each modality. Experiments show that this approach strongly outperforms alternatives that do not exploit the temporal context. The combination of our contributions improve retrieval accuracy up to 10x over prior state of the art. This strong improvement allows us to introduce a wide range of analyses and applications. For instance, we can condition music retrieval based

on visually-defined attributes.
****
Mixed Differential Privacy in Computer Vision
Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8376-8386
We introduce AdaMix, an adaptive differentially private algorithm for training deep neural network classifiers using both private and public image data. While pre-training language models on large public datasets has enabled strong differential privacy (DP) guarantees with minor loss of accuracy, a similar practice yields punishing trade-offs in vision tasks. A few-shot or even zero-shot learning baseline that ignores private data can outperform fine-tuning on a large private dataset. AdaMix incorporates few-shot training, or cross-modal zero-shot learning, on public data prior to private fine-tuning, to improve the trade-off. AdaMix reduces the error increase from the non-private upper bound from the 167-311% of the baseline, on average across 6 datasets, to 68-92% depending on the desired privacy level selected by the user. AdaMix tackles the trade-off arising in visual classification, whereby the most privacy sensitive data, corresponding to isolated points in representation space, are also critical for high classification accuracy. In addition, AdaMix comes with strong theoretical privacy guarantees and convergence analysis.
****
AdaFace: Quality Adaptive Margin for Face Recognition
Minchul Kim, Anil K. Jain, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18750-18759
Recognition in low quality face datasets is challenging because facial attributes are obscured and degraded. Advances in margin-based loss functions have resulted in enhanced discriminability of faces in the embedding space. Further, previous studies have studied the effect of adaptive losses to assign more importance to misclassified (hard) examples. In this work, we introduce another aspect of adaptiveness in the loss function, namely the image quality. We argue that the strategy to emphasize misclassified samples should be adjusted according to their image quality. Specifically, the relative importance of easy or hard samples should be based on the sample's image quality. We propose a new loss function that emphasizes samples of different difficulties based on their image quality. Our method achieves this in the form of an adaptive margin function by approximating the image quality with feature norms. Extensive experiments show that our method, AdaFace, improves the face recognition performance over the state-of-the-art (SoTA) on four datasets (IJB-B, IJB-C, IJB-S and TinyFace). Code and models are released in Supp.
****
Learning Soft Estimator of Keypoint Scale and Orientation With Probabilistic Covariant Loss
Pei Yan, Yihua Tan, Shengzhou Xiong, Yuan Tai, Yansheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19406-19415
Estimating keypoint scale and orientation is crucial to extracting invariant features under significant geometric changes. Recently, the estimators based on self-supervised learning have been designed to adapt to complex imaging conditions. Such learning-based estimators generally predict a single scalar for the keypoint scale or orientation, called hard estimators. However, hard estimators are difficult to handle the local patches containing structures of different objects or multiple edges. In this paper, a Soft Self-Supervised Estimator (S3Esti) is proposed to overcome this problem by learning to predict multiple scales and orientations. S3Esti involves three core factors. First, the estimator is constructed to predict the discrete distributions of scales and orientations. The elements with high confidence will be kept as the final scales and orientations. Second, a probabilistic covariant loss is proposed to improve the consistency of the scale and orientation distributions under different transformations. Third, an optimization algorithm is designed to minimize the loss function, whose convergence

is proved in theory. When combined with different keypoint extraction models, S3 Esti generally improves over 50% accuracy in image matching tasks under signific ant viewpoint changes. In the 3D reconstruction task, S3Esti decreases more than 10% reprojection error and improves the number of registered images.
****

DN-DETR: Accelerate DETR Training by Introducing Query DeNoising
Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 13619-13627
We present in this paper a novel denoising training method to speedup DETR (DEte ction TRansformer) training and offer a deepened understanding of the slow conve rgence issue of DETR-like methods. We show that the slow convergence results fro m the instability of bipartite graph matching which causes inconsistent optimiza tion goals in early training stages. To address this issue, except for the Hunga rian loss, our method additionally feeds ground-truth bounding boxes with noises into Transformer decoder and trains the model to reconstruct the original boxes , which effectively reduces the bipartite graph matching difficulty and leads to a faster convergence. Our method is universal and can be easily plugged into an y DETR-like methods by adding dozens of lines of code to achieve a remarkable im provement. As a result, our DN-DETR results in a remarkable improvement (+1.9AP) under the same setting and achieves the best result (AP 43.4 and 48.6 with 12 a nd 50 epochs of training respectively) among DETR-like methods with ResNet-50 ba ckbone. Our code will be released after the blind review.
****

HCSC: Hierarchical Contrastive Selective Coding
Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, Yi Xu ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 9706-9715
Hierarchical semantic structures naturally exist in an image dataset, in which s everal semantically relevant image clusters can be further integrated into a lar ger cluster with coarser-grained semantics. Capturing such structures with image representations can greatly benefit the semantic understanding on various downs tream tasks. Existing contrastive representation learning methods lack such an i mportant model capability. In addition, the negative pairs used in these methods are not guaranteed to be semantically distinct, which could further hamper the structural correctness of learned image representations. To tackle these limitat ions, we propose a novel contrastive learning framework called Hierarchical Cont rastive Selective Coding (HCSC). In this framework, a set of hierarchical protot ypes are constructed and also dynamically updated to represent the hierarchical semantic structures underlying the data in the latent space. To make image repre sentations better fit such semantic structures, we employ and further improve co nventional instance-wise and prototypical contrastive learning via an elaborate pair selection scheme. This scheme seeks to select more diverse positive pairs w ith similar semantics and more precise negative pairs with truly distinct semant ics. On extensive downstream tasks, we verify the superior performance of HCSC o ver state-of-the-art contrastive methods, and the effectiveness of major model c omponents is proved by plentiful analytical studies. We are continually building a comprehensive model zoo (see supplementary material). Our source code and mod el weights are available at https://github.com/gyfastas/HCSC.
****

TransRank: Self-Supervised Video Representation Learning via Ranking-Based Trans formation Recognition
Haodong Duan, Nanxuan Zhao, Kai Chen, Dahua Lin; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3000-3010
Recognizing transformation types applied to a video clip (RecogTrans) is a long- established paradigm for self-supervised video representation learning, which ac hieves much inferior performance compared to instance discrimination approaches (InstDisc) in recent works. However, based on a thorough comparison of represent ative RecogTrans and InstDisc methods, we observe the great potential of RecogTr ans on both semantic-related and temporal-related downstream tasks. Based on har

d-label classification, existing RecogTrans approaches suffer from noisy supervi
sion signals in pre-training. To mitigate this problem, we developed TransRank,
a unified framework for recognizing Transformations in a Ranking formulation. Tr
ansRank provides accurate supervision signals by recognizing transformations rel
atively, consistently outperforming the classification-based formulation. Meanwh
ile, the unified framework can be instantiated with an arbitrary set of temporal
 or spatial transformations, demonstrating good generality. With a ranking-based
 formulation and several empirical practices, we achieve competitive performance
 on video retrieval and action recognition.Under the same setting, TransRank sur
passes the previous state-of-the-art method by 6.4% on UCF101 and 8.3% on HMDB51
 for action recognition (Top1 Acc); improves video retrieval on UCF101 by 20.4%
(R@1). The promising results validate that RecogTrans is still a worth exploring
 paradigm for video self-supervised learning. Codes will be released at https://
github.com/kennymckormick/TransRank.
****

KeyTr: Keypoint Transporter for 3D Reconstruction of Deformable Objects in Video
s

David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum
, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, Andrea V
edaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2022, pp. 5595-5604

We consider the problem of reconstructing the depth of dynamic objects from vide
os. Recent progress in dynamic video depth prediction has focused on improving t
he output of monocular depth estimators by means of multi-view constraints while
 imposing little to no restrictions on the deformation of the dynamic parts of t
he scene. However, the theory of Non-Rigid Structure from Motion prescribes to c
onstrain the deformations for 3D reconstruction. We thus propose a new model tha
t departs significantly from this prior work. The idea is to fit a dynamic point
 cloud to the video data using Sinkhorn's algorithm to associate the 3D points t
o 2D pixels and use a differentiable point renderer to ensure the compatibility
of the 3D deformations with the measured optical flow. In this manner, our algor
ithm, called Keypoint Transporter, models the overall deformation of the object
within the entire video, so it can constrain the reconstruction correspondingly.
 Compared to weaker deformation models, this significantly reduces the reconstru
ction ambiguity and, for dynamic objects, allows Keypoint Transporter to obtain
reconstructions of the quality superior or at least comparable to prior approach
es while being much faster and reliant on a pre-trained monocular depth estimato
r network. To assess the method, we evaluate on new datasets of synthetic videos
 depicting dynamic humans and animals with ground-truth depth. We also show qual
itative results on crowd-sourced real-world videos of pets.
****

Invariant Grounding for Video Question Answering

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, Tat-Seng Chua; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2
928-2937

Video Question Answering (VideoQA) is the task of answering questions about a vi
deo. At its core is understanding the alignments between visual scenes in video
and linguistic semantics in question to yield the answer. In leading VideoQA mod
els, the typical learning objective, empirical risk minimization (ERM), latches
on superficial correlations between video-question pairs and answers as the alig
nments. However, ERM can be problematic, because it tends to over-exploit the sp
urious correlations between question-irrelevant scenes and answers, instead of i
nspecting the causal effect of question-critical scenes. As a result, the VideoQ
A models suffer from unreliable reasoning. In this work, we first take a causal
look at VideoQA and argue that invariant grounding is the key to ruling out the
spurious correlations. Towards this end, we propose a new learning framework, In
variant Grounding for VideoQA (IGV), to ground the question-critical scene, whos
e causal relations with answers are invariant across different interventions on
the complement. With IGV, the VideoQA models are forced to shield the answering
process from the negative influence of spurious correlations, which significantl

y improves the reasoning ability. Experiments on three benchmark datasets valida te the superiority of IGV in terms of accuracy, visual explainability, and gener alization ability over the leading baselines.
****

Prompt Distribution Learning
Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, Xinmei Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 5206-5215
We present prompt distribution learning for effectively adapting a pre-trained v ision-language model to address downstream recognition tasks. Our method not onl y learns low-bias prompts from a few samples but also captures the distribution of diverse prompts to handle the varying visual representations. In this way, we provide high-quality task-related content for facilitating recognition. This pr ompt distribution learning is realized by an efficient approach that learns the output embeddings of prompts instead of the input embeddings. Thus, we can emplo y a Gaussian distribution to model them effectively and derive a surrogate loss for efficient training. Extensive experiments on 12 datasets demonstrate that ou r method consistently and significantly outperforms existing methods. For exampl e, with 1 sample per category, it relatively improves the average result by 9.1% compared to human-crafted prompts.
****

RAGO: Recurrent Graph Optimizer for Multiple Rotation Averaging
Heng Li, Zhaopeng Cui, Shuaicheng Liu, Ping Tan; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15787-15796
This paper proposes a deep recurrent Rotation Averaging Graph Optimizer (RAGO) f or Multiple Rotation Averaging (MRA). Conventional optimization-based methods us ually fail to produce accurate results due to corrupted and noisy relative measu rements. Recent learning-based approaches regard MRA as a regression problem, wh ile these methods are sensitive to initialization due to the gauge freedom probl em. To handle these problems, we propose a learnable iterative graph optimizer m inimizing a gauge-invariant cost function with an edge rectification strategy to mitigate the effect of inaccurate measurements. Our graph optimizer iteratively refines the global camera rotations by minimizing each node's single rotation o bjective function. Besides, our approach iteratively rectifies relative rotation s to make them more consistent with the current camera orientations and observed relative rotations. Furthermore, we employ a gated recurrent unit to improve th e result by tracing the temporal information of the cost graph. Our framework is a real-time learning-to-optimize rotation averaging graph optimizer with a tiny size deployed for real-world applications. RAGO outperforms previous traditiona l and deep methods on real-world and synthetic datasets. The code is available a t github.com/sfu-gruvi-3dv/RAGO
****

Arch-Graph: Acyclic Architecture Relation Predictor for Task-Transferable Neural Architecture Search
Minbin Huang, Zhijian Huang, Changlin Li, Xin Chen, Hang Xu, Zhenguo Li, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 11881-11891
Neural Architecture Search (NAS) aims to find efficient models for multiple task s. Beyond seeking solutions for a single task, there are surging interests in tr ansferring network design knowledge across multiple tasks. In this line of resea rch, effectively modeling task correlations is vital yet highly neglected. There fore, we propose Arch-Graph, a transferable NAS method that predicts task-specif ic optimal architectures with respect to given task embeddings. It leverages cor relations across multiple tasks by using their embeddings as a part of the predi ctor's input for fast adaptation. We also formulate NAS as an architecture relat ion graph prediction problem, with the relational graph constructed by treating candidate architectures as nodes and their pairwise relations as edges. To enfor ce some basic properties such as acyclicity in the relational graph, we add addi tional constraints to the optimization process, converting NAS into the problem of finding a Maximal Weighted Acyclic Subgraph (MWAS). Our algorithm then strive

s to eliminate cycles and only establish edges in the graph if the rank results can be trusted. Through MWAS, Arch-Graph can effectively rank candidate models for each task with only a small budget to finetune the predictor. With extensive experiments on TransNAS-Bench-101, we show Arch-Graph's transferability and high sample efficiency across numerous tasks, beating many NAS methods designed for both single-task and multi-task search. It is able to find top 0.16% and 0.29% architectures on average on two search spaces under the budget of only 50 models.
****

On Aliased Resizing and Surprising Subtleties in GAN Evaluation
Gaurav Parmar, Richard Zhang, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11410-11420
Metrics for evaluating generative models aim to measure the discrepancy between real and generated images. The oftenused Frechet Inception Distance (FID) metric, for example, extracts "high-level" features using a deep network from the two sets. However, we find that the differences in "low-level" preprocessing, specifically image resizing and compression, can induce large variations and have unforeseen consequences. For instance, when resizing an image, e.g., with a bilinear or bicubic kernel, signal processing principles mandate adjusting prefilter width depending on the downsampling factor, to antialias to the appropriate bandwidth. However, commonly used implementations use a fixed-width prefilter, resulting in aliasing artifacts. Such aliasing leads to corruptions in the feature extraction downstream. Next, lossy compression, such as JPEG, is commonly used to reduce the file size of an image. Although designed to minimally degrade the perceptual quality of an image, the operation also produces variations downstream. Furthermore, we show that if compression is used on real training images, FID can actually improve if the generated images are also subsequently compressed. This paper shows that choices in low-level image processing have been an under-appreciated aspect of generative modeling. We identify and characterize variations in generative modeling development pipelines, provide recommendations based on signal processing principles, and release a reference implementation to facilitate future comparisons.
****

Lepard: Learning Partial Point Cloud Matching in Rigid and Deformable Scenes
Yang Li, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5554-5564
We present Lepard, a Learning based approach for partial point cloud matching in rigid and deformable scenes. The key characteristics are the following techniques that exploit 3D positional knowledge for point cloud matching: 1) An architecture that disentangles point cloud representation into feature space and 3D position space. 2) A position encoding method that explicitly reveals 3D relative distance information through the dot product of vectors. 3) A repositioning technique that modifies the crosspoint-cloud relative positions. Ablation studies demonstrate the effectiveness of the above techniques. In rigid cases, Lepard combined with RANSAC and ICP demonstrates state-of-the-art registration recall of 93.9% / 71.3% on the 3DMatch / 3DLoMatch. In deformable cases, Lepard achieves +27.1% / +34.8% higher non-rigid feature matching recall than the prior art on our newly constructed 4DMatch / 4DLoMatch benchmark.
****

Virtual Elastic Objects
Hsiao-yu Chen, Edith Tretschk, Tuur Stuyck, Petr Kadlecek, Ladislav Kavan, Etienne Vouga, Christoph Lassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15827-15837
We present Virtual Elastic Objects (VEOs): virtual objects that not only look like their real-world counterparts but also behave like them, even when subject to novel interactions. Achieving this presents multiple challenges: not only do objects have to be captured including the physical forces acting on them, then faithfully reconstructed and rendered, but also plausible material parameters found and simulated. To create VEOs, we built a multi-view capture system that captures objects under the influence of a compressed air stream. Building on recent advances in model-free, dynamic Neural Radiance Fields, we reconstruct the objects

and corresponding deformation fields. We propose to use a differentiable, particle-based simulator to use these deformation fields to find representative material parameters, which enable us to run new simulations. To render simulated objects, we devise a method for integrating the simulation results with Neural Radiance Fields. The resulting method is applicable to a wide range of scenarios: it can handle objects composed of inhomogeneous material, with very different shapes, and it can simulate interactions with other virtual objects. We present our results using a newly collected dataset of 12 objects under a variety of force fields, which will be made available upon publication.
****

DiSparse: Disentangled Sparsification for Multitask Model Compression
Xinglong Sun, Ali Hassani, Zhangyang Wang, Gao Huang, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12382-12392
Despite the popularity of Model Compression and Multitask Learning, how to effectively compress a multitask model has been less thoroughly analyzed due to the challenging entanglement of tasks in the parameter space. In this paper, we propose DiSparse, a simple, effective, and first-of-its-kind multitask pruning and sparse training scheme. We consider each task independently by disentangling the importance measurement and take the unanimous decisions among all tasks when performing parameter pruning and selection. Our experimental results demonstrate superior performance on various configurations and settings compared to popular sparse training and pruning methods. Besides the effectiveness in compression, DiSparse also provides a powerful tool to the multitask learning community. Surprisingly, we even observed better performance than some dedicated multitask learning methods in several cases despite the high model sparsity enforced by DiSparse. We analyzed the pruning masks generated with DiSparse and observed strikingly similar sparse network architecture identified by each task even before the training starts. We also observe the existence of a "watershed" layer where the task relatedness sharply drops, implying no benefits in continued parameters sharing. Our code and models will be available at: https://github.com/SHI-Labs/DiSparse-Multitask-Model-Compression.
****

Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference
Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, Timothy M. Hospedales; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9068-9077
Few-shot learning (FSL) is an important and topical problem in computer vision that has motivated extensive research into numerous methods spanning from sophisticated meta-learning methods to simple transfer learning baselines. We seek to push the limits of a simple-but-effective pipeline for real-world few-shot image classification in practice. To this end, we explore few-shot learning from the perspective of neural architecture, as well as a three stage pipeline of pre-training on external data, meta-training with labelled few-shot tasks, and task-specific fine-tuning on unseen tasks. We investigate questions such as: (1) How pre-training on external data benefits FSL? (2) How state of the art transformer architectures can be exploited? and (3) How to best exploit fine-tuning? Ultimately, we show that a simple transformer-based pipeline yields surprisingly good performance on standard benchmarks such as Mini-ImageNet, CIFAR-FS, CDFSL and Meta-Dataset. Our code is available at https://hushell.github.io/pmf.
****

Opening Up Open World Tracking
Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, Laura Leal-Taixé; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19045-19055
Tracking and detecting any object, including ones never-seen-before during model training, is a crucial but elusive capability of autonomous systems. An autonomous agent that is blind to never-seen-before objects poses a safety hazard when operating in the real world - and yet this is how almost all current systems wor

k. One of the main obstacles towards advancing tracking any object is that this task is notoriously difficult to evaluate. A benchmark that would allow us to perform an apple-to-apple comparison of existing efforts is a crucial first step towards advancing this important research field. This paper addresses this evaluation deficit and lays out the landscape and evaluation methodology for detecting and tracking both known and unknown objects in the open-world setting. We propose a new benchmark, TAO-OW: Tracking Any Object in an Open World, analyze existing efforts in multi-object tracking, and construct a baseline for this task while highlighting future challenges. We hope to open a new front in multi-object tracking research that will hopefully bring us a step closer to intelligent systems that can operate safely in the real world.

****

Towards Efficient and Scalable Sharpness-Aware Minimization

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, Yang You; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12360-12370

Recently, Sharpness-Aware Minimization (SAM), which connects the geometry of the loss landscape and generalization, has demonstrated a significant performance boost on training large-scale models such as vision transformers. However, the update rule of SAM requires two sequential (non-parallelizable) gradient computations at each step, which can double the computational overhead. In this paper, we propose a novel algorithm LookSAM - that only periodically calculates the inner gradient ascent, to significantly reduce the additional training cost of SAM. The empirical results illustrate that LookSAM achieves similar accuracy gains to SAM while being tremendously faster - it enjoys comparable computational complexity with first-order optimizers such as SGD or Adam. To further evaluate the performance and scalability of LookSAM, we incorporate a layer-wise modification and perform experiments in the large-batch training scenario, which is more prone to converge to sharp local minima. Equipped with the proposed algorithms, we are the first to successfully scale up the batch size when training Vision Transformers (ViTs). With a 64k batch size, we are able to train ViTs from scratch in minutes while maintaining competitive performance. The code is available here: https://github.com/yong-6/LookSAM

****

VISTA: Boosting 3D Object Detection via Dual Cross-VIew SpaTial Attention

Shengheng Deng, Zhihao Liang, Lin Sun, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8448-8457

Detecting objects from LiDAR point clouds is of tremendous significance in autonomous driving. In spite of good progress, accurate and reliable 3D detection is yet to be achieved due to the sparsity and irregularity of LiDAR point clouds. Among existing strategies, multi-view methods have shown great promise by leveraging the more comprehensive information from both bird's eye view (BEV) and range view (RV). These multi-view methods either refine the proposals predicted from single view via fused features, or fuse the features without considering the global spatial context; their performance is limited consequently. In this paper, we propose to adaptively fuse multi-view features in a global spatial context via Dual Cross-VIew SpaTial Attention (VISTA). The proposed VISTA is a novel plug-and-play fusion module, wherein the multi-layer perceptron widely adopted in standard attention modules is replaced with a convolutional one. Thanks to the learned attention mechanism, VISTA can produce fused features of high quality for prediction of proposals. We decouple the classification and regression tasks in VISTA, and an additional constraint of attention variance is applied that enables the attention module to focus on specific targets instead of generic points. We conduct thorough experiments on the benchmarks of nuScenes and Waymo; results confirm the efficacy of our designs. At the time of submission, our method achieves 63.0% in overall mAP and 69.8% in NDS on the nuScenes benchmark, outperforming all published methods by up to 24% in safety-crucial categories such as cyclist.

****

Rethinking Deep Face Restoration

Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyo

u Chen, Xuhui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7652-7661
A model that can authentically restore a low-quality face image to a high-quality one can benefit many applications. While existing approaches for face restoration make significant progress in generating high-quality faces, they often fail to preserve facial features and cannot authentically reconstruct the faces. Because the human visual system is very sensitive to faces, even minor facial changes may alter the identity and significantly degrade the perceptual quality. In this work, we argue the problems of existing models can be traced down to the two sub-tasks of the face restoration problem, i.e. face generation and face reconstruction, and the fragile balance between them. Based on the observation, we propose a new face restoration model that improves both generation and reconstruction by learning a stochastic model and enhancing the latent features respectively. Furthermore, we adapt the number of skip connections for a better balance between the two sub-tasks. Besides the model improvement, we also introduce a new evaluation metric for measuring models' ability to preserve the identity in the restored faces. Extensive experiments demonstrate that our model achieves state-of-the-art performance on multiple face restoration benchmarks. The user study shows that our model produces higher quality faces while better preserving the identity 86.4% of the time compared with the best performing baselines.
****

OSSO: Obtaining Skeletal Shape From Outside
Marilyn Keller, Silvia Zuffi, Michael J. Black, Sergi Pujades; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20492-20501
We address the problem of inferring the anatomic skeleton of a person, in an arbitrary pose, from the 3D surface of the body; i.e. we predict the inside (bones) from the outside (skin). This has many applications in medicine and biomechanics. Existing state-of-the-art biomechanical skeletons are detailed but do not easily generalize to new subjects. Additionally, computer vision and graphics methods that predict skeletons are typically heuristic, not learned from data, do not leverage the full 3D body surface, and are not validated against ground truth. To our knowledge, our system, called OSSO (Obtaining Skeletal Shape from Outside), is the first to learn the mapping from the 3D body surface to the internal skeleton from real data. We do so using 1000 male and 1000 female dual-energy X-ray absorptiometry (DXA) scans. To these, we fit a parametric 3D body shape model (STAR) to capture the body surface and a novel part-based 3D skeleton model to capture the bones. This provides inside/outside training pairs. We model the statistical variation of full skeletons using PCA in a pose-normalized space and train a regressor from body shape parameters to skeleton shape parameters. Given an arbitrary 3D body shape and pose, OSSO predicts a realistic skeleton inside. In contrast to previous work, we evaluate the accuracy of the skeleton shape quantitatively on held out DXA scans, outperforming the state-of-the art. We also show 3D skeleton prediction from varied and challenging 3D bodies. The code to infer a skeleton from a body shape is available at https://osso.is.tue.mpg.de, and the dataset of paired outer surface (skin) and skeleton (bone) meshes is available as a Biobank Returned Dataset. This research has been conducted using the UK Biobank Resource.
****

Temporal Alignment Networks for Long-Term Video
Tengda Han, Weidi Xie, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2906-2916
The objective of this paper is a temporal alignment network that ingests long term video sequences, and associated text sentences, in order to: (1) determine if a sentence is alignable with the video; and (2) if it is alignable, then determine its alignment. The challenge is to train such networks from large-scale data sets, such as HowTo100M, where the associated text sentences have significant noise, and are only weakly aligned when relevant. Apart from proposing the alignment network, we also make four contributions: (i) we describe a novel co-training method that enables to denoise and train on raw instructional videos without us

ing manual annotation, despite the considerable noise; (ii) to benchmark the alignment performance, we manually curate a 10-hour subset of HowTo100M, totalling 80 videos, with sparse temporal descriptions. Our proposed model, trained on HowTo100M, outperforms strong baselines (CLIP, MIL-NCE) on this alignment dataset by a significant margin; (iii) we apply the trained model in the zero-shot settings to multiple downstream video understanding tasks and achieve state-of-the-art results, including text-video retrieval on YouCook2, and weakly supervised video action segmentation on Breakfast-Action. (iv) we use the automatically-aligned HowTo100M annotations for end-to-end finetuning of the backbone model, and obtain improved performance on downstream action recognition tasks.
****

Few-Shot Head Swapping in the Wild
Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10789-10798
The head swapping task aims at flawlessly placing a source head onto a target body, which is of great importance to various entertainment scenarios. While face swapping has drawn much attention in the community, the task of head swapping has rarely been explored, particularly under the few-shot setting. It is inherently challenging due to its unique needs in head modeling and background blending. In this paper, we present the Head Swapper (HeSer), which achieves few-shot head swapping in the wild through two dedicated designed modules. Firstly, a Head2Head Aligner is devised to holistically migrate position and expression information from the target to the source head by examining multi-scale information. Secondly, to tackle the challenges of skin color variations and head-background mismatches, a Head2Scene Blender is introduced to simultaneously modify facial skin color and fill mismatched gaps on the background around the head. Particularly, seamless blending is achieved through a semantic-guided exemplar warping procedure. User studies and experimental results demonstrate that the proposed method produces superior head swapping results on a variety of scenes.
****

A Study on the Distribution of Social Biases in Self-Supervised Learning Visual Models
Kirill Sirotkin, Pablo Carballeira, Marcos Escudero-Viñolo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10442-10451
Deep neural networks are efficient at learning the data distribution if it is sufficiently sampled. However, they can be strongly biased by non-relevant factors implicitly incorporated in the training data. These include operational biases, such as ineffective or uneven data sampling, but also ethical concerns, as the social biases are implicitly present--even inadvertently, in the training data or explicitly defined in unfair training schedules. In tasks having impact on human processes, the learning of social biases may produce discriminatory, unethical and untrustworthy consequences. It is often assumed that social biases stem from supervised learning on labelled data, and thus, Self-Supervised Learning (SSL) wrongly appears as an efficient and bias-free solution, as it does not require labelled data. However, it was recently proven that a popular SSL method also incorporates biases. In this paper, we study the biases of a varied set of SSL visual models, trained using ImageNet data, using a method and dataset designed by psychological experts to measure social biases. We show that there is a correlation between the type of the SSL model and the number of biases that it incorporates. Furthermore, the results also suggest that this number does not strictly depend on the model's accuracy and changes throughout the network. Finally, we conclude that a careful SSL model selection process can reduce the number of social biases in the deployed model, whilst keeping high performance. The code is available at https://github.com/vpulab/SB-SSL.
****

LAR-SR: A Local Autoregressive Model for Image Super-Resolution
Baisong Guo, Xiaoyun Zhang, Haoning Wu, Yu Wang, Ya Zhang, Yan-Feng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP

R), 2022, pp. 1909-1918
Previous super-resolution (SR) approaches often formulate SR as a regression problem and pixel wise restoration, which leads to a blurry and unreal SR output. Recent works combine adversarial loss with pixel-wise loss to train a GAN-based model or introduce normalizing flows into SR problems to generate more realistic images. As another powerful generative approach, autoregressive (AR) model has not been noticed in low level tasks due to its limitation. Based on the fact that given the structural information, the textural details in the natural images are locally related without long term dependency, in this paper we propose a novel autoregressive model-based SR approach, namely LAR-SR, which can efficiently generate realistic SR images using a novel local autoregressive (LAR) module. The proposed LAR module can sample all the patches of textural components in parallel, which greatly reduces the time consumption. In addition to high time efficiency, it is also able to leverage contextual information of pixels and can be optimized with a consistent loss. Experimental results on the widely-used datasets show that the proposed LAR-SR approach achieves superior performance on the visual quality and quantitative metrics compared with other generative models such as GAN, Flow, and is competitive with the mixture generative model.
****

Bayesian Invariant Risk Minimization
Yong Lin, Hanze Dong, Hao Wang, Tong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16021-16030
Generalization under distributional shift is an open challenge for machine learning. Invariant Risk Minimization (IRM) is a promising framework to tackle this issue by extracting invariant features. However, despite the potential and popularity of IRM, recent works have reported negative results of it on deep models. We argue that the failure can be primarily attributed to deep models' tendency to overfit the data. Specifically, our theoretical analysis shows that IRM degenerates to empirical risk minimization (ERM) when overfitting occurs. Our empirical evidence also provides supports: IRM methods that work well in typical settings significantly deteriorate even if we slightly enlarge the model size or lessen the training data. To alleviate this issue, we propose Bayesian Invariant Risk Minimization (BIRM) by introducing Bayesian inference into the IRM. The key motivation is to estimate the penalty of IRM based on the posterior distribution of classifiers (as opposed to a single classifier), which is much less prone to overfitting. Extensive experimental results on four datasets demonstrate that BIRM consistently outperforms the existing IRM baselines significantly.
****

Democracy Does Matter: Comprehensive Feature Mining for Co-Salient Object Detection
Siyue Yu, Jimin Xiao, Bingfeng Zhang, Eng Gee Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 979-988
Co-salient object detection, with the target of detecting co-existed salient objects among a group of images, is gaining popularity. Recent works use the attention mechanism or extra information to aggregate common co-salient features, leading to incomplete even incorrect responses for target objects. In this paper, we aim to mine comprehensive co-salient features with democracy and reduce background interference without introducing any extra information. To achieve this, we design a democratic prototype generation module to generate democratic response maps, covering sufficient co-salient regions and thereby involving more shared attributes of co-salient objects. Then a comprehensive prototype based on the response maps can be generated as a guide for final prediction. To suppress the noisy background information in the prototype, we propose a self-contrastive learning module, where both positive and negative pairs are formed without relying on additional classification information. Besides, we also design a democratic feature enhancement module to further strengthen the co-salient features by readjusting attention values. Extensive experiments show that our model obtains better performance than previous state-of-the-art methods, especially on challenging real-world cases (e.g., for CoCA, we obtain a gain of 2.0% for MAE, 5.4% for maximum F-measure, 2.3% for maximum E-measure, and 3.7% for S-measure) under the same

settings. Code will be released soon.
****
Alleviating Semantics Distortion in Unsupervised Low-Level Image-to-Image Translation via Structure Consistency Constraint
Jiaxian Guo, Jiachen Li, Huan Fu, Mingming Gong, Kun Zhang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18249-18259
Unsupervised image-to-image (I2I) translation aims to learn a domain mapping function that can preserve the semantics of the input images without paired data. However, because the underlying semantics distributions in the source and target domains are often mismatched, current distribution matching-based methods may distort the semantics when matching distributions, resulting in the inconsistency between the input and translated images, which is known as the semantics distortion problem. In this paper, we focus on the low-level I2I translation, where the structure of images is highly related to their semantics. To alleviate semantic distortions in such translation tasks without paired supervision, we propose a novel I2I translation constraint, called Structure Consistency Constraint (SCC), to promote the consistency of image structures by reducing the randomness of color transformation in the translation process. To facilitate estimation and maximization of SCC, we propose an approximate representation of mutual information called relative Squared-loss Mutual Information (rSMI) that enjoys efficient analytic solutions. Our SCC can be easily incorporated into most existing translation models. Quantitative and qualitative comparisons on a range of low-level I2I translation tasks show that translation models with SCC outperform the original models by a significant margin with little additional computational and memory costs.
****
Doodle It Yourself: Class Incremental Learning by Drawing a Few Sketches
Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2293-2302
The human visual system is remarkable in learning new visual concepts from just a few examples. This is precisely the goal behind few-shot class incremental learning (FSCIL), where the emphasis is additionally placed on ensuring the model does not suffer from "forgetting". In this paper, we push the boundary further for FSCIL by addressing two key questions that bottleneck its ubiquitous application (i) can the model learn from diverse modalities other than just photo (as humans do), and (ii) what if photos are not readily accessible (due to ethical and privacy constraints). Our key innovation lies in advocating the use of sketches as a new modality for class support. The product is a "Doodle It Yourself" (DIY) FSCIL framework where the users can freely sketch a few examples of a novel class for the model to learn to recognise photos of that class. For that, we present a framework that infuses (i) gradient consensus for domain invariant learning, (ii) knowledge distillation for preserving old class information, and (iii) graph attention networks for message passing between old and novel classes. We experimentally show that sketches are better class support than text in the context of FSCIL, echoing findings elsewhere in the sketching literature.
****
Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes
Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3222-3231
Sound source localization in visual scenes aims to localize objects emitting the sound in a given image. Recent works showing impressive localization performance typically rely on the contrastive learning framework. However, the random sampling of negatives, as commonly adopted in these methods, can result in misalignment between audio and visual features and thus inducing ambiguity in localization. In this paper, instead of following previous literature, we propose Self-Supervised Predictive Learning (SSPL), a negative-free method for sound localization

via explicit positive mining. Specifically, we first devise a three-stream network to elegantly associate sound source with two augmented views of one corresponding video frame, leading to semantically coherent similarities between audio and visual features. Second, we introduce a novel predictive coding module for audio-visual feature alignment. Such a module assists SSPL to focus on target objects in a progressive manner and effectively lowers the positive-pair learning difficulty. Experiments show surprising results that SSPL outperforms the state-of-the-art approach on two standard sound localization benchmarks. In particular, SSPL achieves significant improvements of 8.6% cIoU and 3.4% AUC on SoundNet-Flickr compared to the previous best. Code is available at: https://github.com/zjsong/SSPL.

****

ICON: Implicit Clothed Humans Obtained From Normals

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13296-13306

Current methods for learning realistic and animatable 3D clothed avatars need either posed 3D scans or 2D images with carefully controlled user poses. In contrast, our goal is to learn the avatar from only 2D images of people in unconstrained poses. Given a set of images, our method estimates a detailed 3D surface from each image and then combines these into an animatable avatar. Implicit functions are well suited to the first task, as they can capture details like hair or clothes. Current methods, however, are not robust to varied human poses and often produce 3D surfaces with broken or disembodied limbs, missing details, or non-human shapes. The problem is that these methods use global feature encoders that are sensitive to global pose. To address this, we propose ICON ("Implicit Clothed humans Obtained from Normals"), which uses local features. ICON has two main modules, both of which exploit the SMPL body model. First, ICON infers detailed clothed-human normals(front/back) conditioned on the SMPL normals. Second, a visibility-aware implicit surface regressor produces an iso-surface of the human occupancy field. Importantly, at inference time, a feedback loop alternates between refining the SMPL mesh using the inferred clothed normals and then refining the normals. Given multiple reconstructed frames of a subject in varied poses, we use modified SCANimate to produce an animatable avatar from them. Evaluation on the AGORA and CAPE datasets shows that ICON outperforms the state-of-the-art in reconstruction, even with heavily limited training data. Additionally, it is much more robust to out-of-distribution samples, e.g., in-the-wild poses/images and out-of-frame cropping. ICON takes a step towards pose-robust 3D clothed human reconstruction from in-the-wild images. This enables creating avatars directly from video with personalized and nature pose-dependent cloth deformation. Our models and code will be available for research.

****

Comparing Correspondences: Video Prediction With Correspondence-Wise Losses

Daniel Geng, Max Hamilton, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3365-3376

Image prediction methods often struggle on tasks that require changing the positions of objects, such as video prediction, producing blurry images that average over the many positions that objects might occupy. In this paper, we propose a simple change to existing image similarity metrics that makes them more robust to positional errors: we match the images using optical flow, then measure the visual similarity of corresponding pixels. This change leads to crisper and more perceptually accurate predictions, and does not require modifications to the image prediction network. We apply our method to a variety of video prediction tasks, where it obtains strong performance with simple network architectures, and to the closely related task of video interpolation. Code and results are available at our webpage: https://dangeng.github.io/CorrWiseLosses

****

Uni-Perceiver: Pre-Training Unified Architecture for Generic Perception for Zero-Shot and Few-Shot Tasks

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, Jifeng D

ai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2022, pp. 16804-16815
Biological intelligence systems of animals perceive the world by integrating inf
ormation in different modalities and processing simultaneously for various tasks
. In contrast, current machine learning research follows a task-specific paradig
m, leading to inefficient collaboration between tasks and high marginal costs of
 developing perception models for new tasks. In this paper, we present a generic
 perception architecture named Uni-Perceiver, which processes a variety of modal
ities and tasks with unified modeling and shared parameters. Specifically, Uni-P
erceiver encodes different task inputs and targets from arbitrary modalities int
o a unified representation space with a modality-agnostic Transformer encoder an
d lightweight modality-specific tokenizers. Different perception tasks are model
ed as the same formulation, that is, finding the maximum likelihood target for e
ach input through the similarity of their representations. The model is pre-trai
ned on several uni-modal and multi-modal tasks, and evaluated on a variety of do
wnstream tasks, including novel tasks that did not appear in the pre-training st
age. Results show that our pre-trained model without any tuning can achieve reas
onable performance even on novel tasks. The performance can be improved to a lev
el close to state-of-the-art methods by conducting prompt tuning on 1% of downst
ream task data. Full-data fine-tuning further delivers results on par with or be
tter than state-of-the-art results. Code and pre-trained weights shall be releas
ed.
****

The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Mo
nitoring Under Domain Shift
Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee M
ukherjee, Stanley Chan, John Morgan, Vivek Rathod, Jonathan Huang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
2, pp. 21294-21307
Generalization to novel domains is a fundamental challenge for computer vision.
Near-perfect accuracy on benchmarks is common, but these models do not work as e
xpected when deployed outside of the training distribution. To build computer vi
sion systems that truly solve real-world problems at global scale, we need bench
marks that fully capture real-world complexity, including geographic domain shif
t, long-tailed distributions, and data noise. We propose urban forest monitoring
 as an ideal testbed for studying and improving upon these computer vision chall
enges, while simultaneously working towards filling a crucial environmental and
societal need. Urban forests provide significant benefits to urban societies (e.
g., cleaner air and water, carbon sequestration, and energy savings among others
). However, planning and maintaining these forests is expensive. One particularl
y costly aspect of urban forest management is monitoring the existing trees in a
 city: e.g., tracking tree locations, species, and health. Monitoring efforts ar
e currently based on tree censuses built by human experts, costing cities millio
ns of dollars per census and thus collected infrequently. Previous investigation
s into automating urban forest monitoring focused on small datasets from single
cities, covering only common categories. To address these shortcomings, we intro
duce a new large-scale dataset that joins public tree censuses from 23 cities wi
th a large collection of street level and aerial imagery. Our Auto Arborist data
set contains over 2.5M trees and 344 genera and is >2 orders of magnitude larger
 than the closest dataset in the literature. We introduce baseline results on ou
r dataset across modalities as well as metrics for the detailed analysis of gene
ralization with respect to geographic distribution shifts, vital for such a syst
em to be deployed at-scale.
****

On the Instability of Relative Pose Estimation and RANSAC's Role
Hongyi Fan, Joe Kileel, Benjamin Kimia; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8935-8943
Relative pose estimation using the 5-point or 7-point Random Sample Consensus (R
ANSAC) algorithms can fail even when no outliers are present and there are enoug
h inliers to support a hypothesis. These cases arise due to numerical instabilit

y of the 5- and 7-point minimal problems. This paper characterizes these instabi
lities, both in terms of minimal world scene configurations that lead to infinit
e condition number in epipolar estimation, and also in terms of the related mini
mal image feature pair correspondence configurations. The instability is studied
 in the context of a novel framework for analyzing the conditioning of minimal p
roblems in multiview geometry, based on Riemannian manifolds. Experiments with s
ynthetic and real-world data reveal that RANSAC does not only serve to filter ou
t outliers, but RANSAC also selects for well-conditioned image data, sufficientl
y separated from the ill-posed locus that our theory predicts. These findings su
ggest that, in future work, one could try to accelerate and increase the success
 of RANSAC by testing only well-conditioned image data.
****

Shape From Polarization for Complex Scenes in the Wild
Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, Qifeng Chen; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 12632-12641
We present a new data-driven approach with physics-based priors to scene-level n
ormal estimation from a single polarization image. Existing shape from polarizat
ion (SfP) works mainly focus on estimating the normal of a single object rather
 than complex scenes in the wild. A key barrier to high-quality scene-level SfP i
s the lack of real-world SfP data in complex scenes. Hence, we contribute the fi
rst real-world scene-level SfP dataset with paired input polarization images and
 ground-truth normal maps. Then we propose a learning-based framework with a mul
ti-head self-attention module and viewing encoding, which is designed to handle
 increasing polarization ambiguities caused by complex materials and non-orthogra
phic projection in scene-level SfP. Our trained model can be generalized to far-
field outdoor scenes as the relationship between polarized light and surface nor
mals is not affected by distance. Experimental results demonstrate that our appr
oach significantly outperforms existing SfP models on two datasets. Our dataset
 and source code will be publicly available.
****

Real-Time, Accurate, and Consistent Video Semantic Segmentation via Unsupervised
 Adaptation and Cross-Unit Deployment on Mobile Device
Hyojin Park, Alan Yessenbayev, Tushar Singhal, Navin Kumar Adhikari, Yizhe Zhang
, Shubhankar Mangesh Borse, Hong Cai, Nilesh Prasad Pandey, Fei Yin, Frank Mayer
, Balaji Calidas, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2022, pp. 21431-21438
This demonstration showcases our innovations on efficient, accurate, and tempora
lly consistent video semantic segmentation on mobile device. We employ our test-
time unsupervised scheme, AuxAdapt, to enable the segmentation model to adapt to
 a given video in an online manner. More specifically, we leverage a small auxil
iary network to perform weight updates and keep the large, main segmentation net
work frozen. This significantly reduces the computational cost of adaptation whe
n compared to previous methods (e.g., Tent, DVP), and at the same time, prevents
 catastrophic forgetting. By running AuxAdapt, we can considerably improve the t
emporal consistency of video segmentation while maintaining the accuracy. We dem
onstrate how to efficiently deploy our adaptive video segmentation algorithm on
a smartphone powered by a Snapdragon Mobile Platform. Rather than simply running
 the entire algorithm on the GPU, we adopt a cross-unit deployment strategy. The
 main network, which will be frozen during test time, will perform inferences on
 a highly optimized AI accelerator unit, while the small auxiliary network, whic
h will be updated on the fly, will run forward passes and back-propagations on t
he GPU. Such a deployment scheme best utilizes the available processing power on
 the smartphone and enables real-time operation of our adaptive video segmentati
on algorithm. We provide example videos in supplementary material.
****

SNUG: Self-Supervised Neural Dynamic Garments
Igor Santesteban, Miguel A. Otaduy, Dan Casas; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8140-8150
We present a self-supervised method to learn dynamic 3D deformations of garments

worn by parametric human bodies. State-of-the-art data-driven approaches to model 3D garment deformations are trained using supervised strategies that require large datasets, usually obtained by expensive physics-based simulation methods or professional multi-camera capture setups. In contrast, we propose a new training scheme that removes the need for ground-truth samples, enabling self-supervised training of dynamic 3D garment deformations. Our key contribution is to realize that physics-based deformation models, traditionally solved in a frame-by-frame basis by implicit integrators, can be recasted as an optimization problem. We leverage such optimization-based scheme to formulate a set of physics-based loss terms that can be used to train neural networks without precomputing ground-truth data. This allows us to learn models for interactive garments, including dynamic deformations and fine wrinkles, with two orders of magnitude speed up in training time compared to state-of-the-art supervised methods.
****

Towards Fewer Annotations: Active Learning via Region Impurity and Prediction Uncertainty for Domain Adaptive Semantic Segmentation
Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8068-8078
Self-training has greatly facilitated domain adaptive semantic segmentation, which iteratively generates pseudo labels on unlabeled target data and retrains the network. However, realistic segmentation datasets are highly imbalanced, pseudo labels are typically biased to the majority classes and basically noisy, leading to an error-prone and suboptimal model. In this paper, we propose a simple region-based active learning approach for semantic segmentation under a domain shift, aiming to automatically query a small partition of image regions to be labeled while maximizing segmentation performance. Our algorithm, Region Impurity and Prediction Uncertainty (RIPU), introduces a new acquisition strategy characterizing the spatial adjacency of image regions along with the prediction confidence. We show that the proposed region-based selection strategy makes more efficient use of a limited budget than image-based or point-based counterparts. Further, we enforce local prediction consistency between a pixel and its nearest neighbors on a source image. Alongside, we develop a negative learning loss to make the features more discriminative. Extensive experiments demonstrate that our method only requires very few annotations to almost reach the supervised performance and substantially outperforms state-of-the-art methods. The code is available at https://github.com/BIT-DA/RIPU.
****

Glass Segmentation Using Intensity and Spectral Polarization Cues
Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12622-12631
Transparent and semi-transparent materials pose significant challenges for existing scene understanding and segmentation algorithms due to their lack of RGB texture which impedes the extraction of meaningful features. In this work, we exploit that the light-matter interactions on glass materials provide unique intensity-polarization cues for each observed wavelength of light. We present a novel learning-based glass segmentation network that leverages both trichromatic (RGB) intensities as well as trichromatic linear polarization cues from a single photograph captured without making any assumption on the polarization state of the illumination. Our novel network architecture dynamically fuses and weights both the trichromatic color and polarization cues using a novel global-guidance and multi-scale self-attention module, and leverages global cross-domain contextual information to achieve robust segmentation. We train and extensively validate our segmentation method on a new large-scale RGB-Polarization dataset (RGBP-Glass), and demonstrate that our method outperforms state-of-the-art segmentation approaches by a significant margin.
****

CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, Ranga Rodrigo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9902-9912
Manual annotation of large-scale point cloud dataset for varying tasks such as 3D object classification, segmentation and detection is often laborious owing to the irregular structure of point clouds. Self-supervised learning, which operates without any human labeling, is a promising approach to address this issue. We observe in the real world that humans are capable of mapping the visual concepts learnt from 2D images to understand the 3D world. Encouraged by this insight, we propose CrossPoint, a simple cross-modal contrastive learning approach to learn transferable 3D point cloud representations. It enables a 3D-2D correspondence of objects by maximizing agreement between point clouds and the corresponding rendered 2D image in the invariant space, while encouraging invariance to transformations in the point cloud modality. Our joint training objective combines the feature correspondences within and across modalities, thus ensembles a rich learning signal from both 3D point cloud and 2D image modalities in a self-supervised fashion. Experimental results show that our approach outperforms the previous unsupervised learning methods on a diverse range of downstream tasks including 3D object classification and segmentation. Further, the ablation studies validates the potency of our approach for a better point cloud understanding. Code and pretrained models are available at https://github.com/MohamedAfham/CrossPoint.
****

Few Shot Generative Model Adaption via Relaxed Spatial Structural Alignment
Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11204-11213
Training a generative adversarial network (GAN) with limited data has been a challenging task. A feasible solution is to start with a GAN well-trained on a large scale source domain and adapt it to the target domain with a few samples, termed as few shot generative model adaption. However, existing methods are prone to model overfitting and collapse in extremely few shot setting (less than 10). To solve this problem, we propose a relaxed spatial structural alignment (RSSA) method to calibrate the target generative models during the adaption. We design a cross-domain spatial structural consistency loss comprising the self-correlation and disturbance correlation consistency loss. It helps align the spatial structural information between the synthesis image pairs of the source and target domains. To relax the cross-domain alignment, we compress the original latent space of generative models to a subspace. Image pairs generated from the subspace are pulled closer. Qualitative and quantitative experiments show that our method consistently surpasses the state-of-the-art methods in few shot setting. Our source code: https://github.com/StevenShaw1999/RSSA.
****

Target-Relevant Knowledge Preservation for Multi-Source Domain Adaptive Object Detection
Jiaxi Wu, Jiaxin Chen, Mengzhe He, Yiru Wang, Bo Li, Bingqi Ma, Weihao Gan, Wei Wu, Yali Wang, Di Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5301-5310
Domain adaptive object detection (DAOD) is a promising way to alleviate performance drop of detectors in new scenes. Albeit great effort made in single source domain adaptation, a more generalized task with multiple source domains remains not being well explored, due to knowledge degradation during their combination. To address this issue, we propose a novel approach, namely target-relevant knowledge preservation (TRKP), to unsupervised multi-source DAOD. Specifically, TRKP adopts the teacher-student framework, where the multi-head teacher network is built to extract knowledge from labeled source domains and guide the student network to learn detectors in unlabeled target domain. The teacher network is further equipped with an adversarial multi-source disentanglement (AMSD) module to preserve source domain-specific knowledge and simultaneously perform cross-domain alignment. Besides, a holistic target-relevant mining (HTRM) scheme is developed to re-weight the source images according to the source-target relevance. By this m

eans, the teacher network is enforced to capture target-relevant knowledge, thus benefiting decreasing domain shift when mentoring object detection in the target domain. Extensive experiments are conducted on various widely used benchmarks with new state-of-the-art scores reported, highlighting the effectiveness.
****

Pyramid Grafting Network for One-Stage High Resolution Saliency Detection
Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, Jia Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11717-11726
Recent salient object detection (SOD) methods based on deep neural network have achieved remarkable performance. However, most of existing SOD models designed for low-resolution input perform poorly on high-resolution images due to the contradiction between the sampling depth and the receptive field size. Aiming at resolving this contradiction, we propose a novel one-stage framework called Pyramid Grafting Network (PGNet), using transformer and CNN backbone to extract features from different resolution images independently and then graft the features from transformer branch to CNN branch. An attention-based Cross-Model Grafting Module (CMGM) is proposed to enable CNN branch to combine broken detailed information more holistically, guided by different source feature during decoding process. Moreover, we design an Attention Guided Loss (AGL) to explicitly supervise the attention matrix generated by CMGM to help the network better interact with the attention from different models. We contribute a new Ultra-High-Resolution Saliency Detection dataset UHRSD, containing 5,920 images at 4K-8K resolutions. To our knowledge, it is the largest dataset in both quantity and resolution for high-resolution SOD task, which can be used for training and testing in future research. Sufficient experiments on UHRSD and widely-used SOD datasets demonstrate that our method achieves superior performance compared to the state-of-the-art methods.
****

A Style-Aware Discriminator for Controllable Image Translation
Kunhee Kim, Sanghun Park, Eunyeong Jeon, Taehun Kim, Daijin Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18239-18248
Current image-to-image translations do not control the output domain beyond the classes used during training, nor do they interpolate between different domains well, leading to implausible results. This limitation largely arises because labels do not consider the semantic distance. To mitigate such problems, we propose a style-aware discriminator that acts as a critic as well as a style encoder to provide conditions. The style-aware discriminator learns a controllable style space using prototype-based self-supervised learning and simultaneously guides the generator. Experiments on multiple datasets verify that the proposed model outperforms current state-of-the-art image-to-image translation methods. In contrast with current methods, the proposed approach supports various applications, including style interpolation, content transplantation, and local image translation. The code is available at github.com/kunheek/style-aware-discriminator.
****

Non-Iterative Recovery From Nonlinear Observations Using Generative Models
Jiulong Liu, Zhaoqiang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 233-243
In this paper, we aim to estimate the direction of an underlying signal from its nonlinear observations following the semi-parametric single index model (SIM). Unlike for conventional compressed sensing where the signal is assumed to be sparse, we assume that the signal lies in the range of an L-Lipschitz continuous generative model with bounded k-dimensional inputs. This is mainly motivated by the tremendous success of deep generative models in various real applications. Our reconstruction method is non-iterative (though approximating the projection step may require an iterative procedure) and highly efficient, and it is shown to attain the near-optimal statistical rate of order \sqrt (k \log L)/m , where m is the number of measurements. We consider two specific instances of the SIM, namely noisy 1-bit and cubic measurement models, and perform experiments on image da

tasets to demonstrate the efficacy of our method. In particular, for the noisy 1-bit measurement model, we show that our non-iterative method significantly outperforms a state-of-the-art iterative method in terms of both accuracy and efficiency.
****

Incremental Cross-View Mutual Distillation for Self-Supervised Medical CT Synthesis
Chaowei Fang, Liang Wang, Dingwen Zhang, Jun Xu, Yixuan Yuan, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20677-20686
Due to the constraints of the imaging device and high cost in operation time, computer tomography (CT) scans are usually acquired with low within-slice resolution. Improving the inter-slice resolution is beneficial to the disease diagnosis for both human experts and computer-aided systems. To this end, this paper builds a novel medical slice synthesis to increase the inter-slice resolution. Considering that the ground-truth intermediate medical slices are always absent in clinical practice, we introduce the incremental cross-view mutual distillation strategy to accomplish this task in the self-supervised learning manner. Specifically, we model this problem from three different views: slice-wise interpolation from axial view and pixel-wise interpolation from coronal and sagittal views. Under this circumstance, the models learned from different views can distill valuable knowledge to guide the learning processes of each other. We can repeat this process to make the models synthesize intermediate slice data with increasing between-slice resolution. To demonstrate the effectiveness of the proposed approach, we conduct comprehensive experiments on a large-scale CT dataset. Quantitative and qualitative comparison results show that our method outperforms state-of-the-art algorithms by clear margins.
****

Enhancing Adversarial Training With Second-Order Statistics of Weights
Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, Xiaowei Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15273-15283
Adversarial training has been shown to be one of the most effective approaches to improve the robustness of deep neural networks. It is formalized as a min-max optimization over model weights and adversarial perturbations, where the weights can be optimized through gradient descent methods like SGD. In this paper, we show that treating model weights as random variables allows for enhancing adversarial training through Second-Order Statistics Optimization (S^2O) with respect to the weights. By relaxing a common (but unrealistic) assumption of previous PAC-Bayesian frameworks that all weights are statistically independent, we derive an improved PAC-Bayesian adversarial generalization bound, which suggests that optimizing second-order statistics of weights can effectively tighten the bound. In addition to this theoretical insight, we conduct an extensive set of experiments, which show that S^2O not only improves the robustness and generalization of the trained neural networks when used in isolation, but also integrates easily in state-of-the-art adversarial training techniques like TRADES, AWP, MART, and AVMixup, leading to a measurable improvement of these techniques. The code is available at https://github.com/Alexkael/S2O.
****

Partially Does It: Towards Scene-Level FG-SBIR With Partial Input
Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2395-2405
We scrutinise an important observation plaguing scene-level sketch research -- that a significant portion of scene sketches are "partial". A quick pilot study reveals: (i) a scene sketch does not necessarily contain all objects in the corresponding photo, due to the subjective holistic interpretation of scenes, (ii) there exists significant empty (white) regions as a result of object-level abstraction, and as a result, (iii) existing scene-level fine-grained sketch-based image retrieval methods collapse as scene sketches become more partial. To solve thi

s "partial" problem, we advocate for a simple set-based approach using optimal t
ransport (OT) to model cross-modal region associativity in a partially-aware fas
hion. Importantly, we improve upon OT to further account for holistic partialnes
s by comparing intra-modal adjacency matrices. Our proposed method is not only r
obust to partial scene-sketches but also yields state-of-the-art performance on
existing datasets.
****

Dual Temperature Helps Contrastive Learning Without Many Negative Samples: Towar
ds Understanding and Simplifying MoCo
Chaoning Zhang, Kang Zhang, Trung X. Pham, Axi Niu, Zhinan Qiao, Chang D. Yoo, I
n So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2022, pp. 14441-14450
Contrastive learning (CL) is widely known to require many negative samples, 6553
6 in MoCo for instance, for which the performance of a dictionary-free framework
 is often inferior because the negative sample size (NSS) is limited by its mini
-batch size (MBS). To decouple the NSS from the MBS, a dynamic dictionary has be
en adopted in a large volume of CL frameworks, among which arguably the most pop
ular one is MoCo family. In essence, MoCo adopts a momentum-based queue dictiona
ry, for which we perform a fine-grained analysis of its size and consistency. We
 point out that InfoNCE loss used in MoCo implicitly attract anchors to their co
rresponding positive sample with various strength of penalties and identify such
 inter-anchor hardness-awareness property as a major reason for the necessity of
 a large dictionary. Our findings motivate us to simplify MoCo v2 via the remova
l of its dictionary as well as momentum. Based on an InfoNCE with the proposed d
ual temperature, our simplified frameworks, SimMoCo and SimCo, outperform MoCo v
2 by a visible margin. Moreover, our work bridges the gap between CL and non-CL
frameworks, contributing to a more unified understanding of these two mainstream
 frameworks in SSL. Code is available at: https://bit.ly/3LkQbaT.
****

Moving Window Regression: A Novel Approach to Ordinal Regression
Nyeong-Ho Shin, Seon-Ho Lee, Chang-Su Kim; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18760-18769
A novel ordinal regression algorithm, called moving window regression (MWR), is
proposed in this paper. First, we propose the notion of relative rank (rho-rank)
, which is a new order representation scheme for input and reference instances.
Second, we develop global and local relative regressors (rho-regressors) to pred
ict rho-ranks within entire and specific rank ranges, respectively. Third, we re
fine an initial rank estimate iteratively by selecting two reference instances t
o form a search window and then estimating the rho-rank within the window. Exten
sive experiments results show that the proposed algorithm achieves the state-of-
the-art performances on various benchmark datasets for facial age estimation and
 historical color image classification. The codes are available at https://githu
b.com/nhshin-mcl/MWR.
****

UniCoRN: A Unified Conditional Image Repainting Network
Jimeng Sun, Shuchen Weng, Zheng Chang, Si Li, Boxin Shi; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 113
69-11378
Conditional image repainting (CIR) is an advanced image editing task, which requ
ires the model to generate visual content in user-specified regions conditioned
on multiple cross-modality constraints, and composite the visual content with th
e provided background seamlessly. Existing methods based on two-phase architectu
re design assume dependency between phases and cause color-image incongruity. To
 solve these problems, we propose a novel Unified Conditional image Repainting N
etwork (UniCoRN). We break the two-phase assumption in CIR task by constructing
the interaction and dependency relationship between background and other conditi
ons. We further introduce the hierarchical structure into cross-modality similar
ity model to capture feature patterns at different levels and bridge the gap bet
ween visual content and color condition. A new LANDSCAPE-CIR dataset is collecte
d and annotated to expand the application scenarios of the CIR task. Experiments

show that UniCoRN achieves higher synthetic quality, better condition consistency, and more realistic compositing effect.
****

Forecasting Characteristic 3D Poses of Human Actions
Christian Diller, Thomas Funkhouser, Angela Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15914-15923
We propose the task of forecasting characteristic 3d poses: from a short sequence observation of a person, predict a future 3d pose of that person in a likely action-defining, characteristic pose - for instance, from observing a person picking up an apple, predict the pose of the person eating the apple. Prior work on human motion prediction estimates future poses at fixed time intervals. Although easy to define, this frame-by-frame formulation confounds temporal and intentional aspects of human action. Instead, we define a semantically meaningful pose prediction task that decouples the predicted pose from time, taking inspiration from goal-directed behavior. To predict characteristic poses, we propose a probabilistic approach that models the possible multi-modality in the distribution of likely characteristic poses. We then sample future pose hypotheses from the predicted distribution in an autoregressive fashion to model dependencies between joints. To evaluate our method, we construct a dataset of manually annotated characteristic 3d poses. Our experiments with this dataset suggest that our proposed probabilistic approach outperforms state-of-the-art methods by 26% on average.
****

ACPL: Anti-Curriculum Pseudo-Labelling for Semi-Supervised Medical Image Classification
Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, Gustavo Carneiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20697-20706
Effective semi-supervised learning (SSL) in medical image analysis (MIA) must address two challenges: 1) work effectively on both multi-class (e.g., lesion classification) and multi-label (e.g., multiple-disease diagnosis) problems, and 2) handle imbalanced learning (because of the high variance in disease prevalence). One strategy to explore in SSL MIA is based on the pseudo labelling strategy, but it has a few shortcomings. Pseudo-labelling has in general lower accuracy than consistency learning, it is not specifically designed for both multi-class and multi-label problems, and it can be challenged by imbalanced learning. In this paper, unlike traditional methods that select confident pseudo label by threshold, we propose a new SSL algorithm, called anti-curriculum pseudo-labelling (ACPL), which introduces novel techniques to select informative unlabelled samples, improving training balance and allowing the model to work for both multi-label and multi-class problems, and to estimate pseudo labels by an accurate ensemble of classifiers (improving pseudo label accuracy). We run extensive experiments to evaluate ACPL on two public medical image classification benchmarks: Chest X-Ray 14 for thorax disease multi-label classification and ISIC2018 for skin lesion multi-class classification. Our method outperforms previous SOTA SSL methods on both datasets
****

Learning to Deblur Using Light Field Generated and Real Defocus Images
Lingyan Ruan, Bin Chen, Jizhou Li, Miuling Lam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16304-16313
Defocus deblurring is a challenging task due to the spatially varying nature of defocus blur. While deep learning approach shows great promise in solving image restoration problems, defocus deblurring demands accurate training data that consists of all-in-focus and defocus image pairs, which is difficult to collect. Naive two-shot capturing cannot achieve pixel-wise correspondence between the defocused and all-in-focus image pairs. Synthetic aperture of light fields is suggested to be a more reliable way to generate accurate image pairs. However, the defocus blur generated from light field data is different from that of the images captured with a traditional digital camera. In this paper, we propose a novel deep defocus deblurring network that leverages the strength and overcomes the shortcoming of light fields. We first train the network on a light field-generated da

taset for its highly accurate image correspondence. Then, we fine-tune the netwo rk using feature loss on another dataset collected by the two-shot method to all eviate the differences between the defocus blur exists in the two domains. This strategy is proved to be highly effective and able to achieve the state-of-the-a rt performance both quantitatively and qualitatively on multiple test sets. Exte nsive ablation studies have been conducted to analyze the effect of each network module to the final performance.
****

Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection
Nicolae-C■t■lin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, Mubarak Shah; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13576-13586
Anomaly detection is commonly pursued as a one-class classification problem, whe re models can only learn from normal training samples, while being evaluated on both normal and abnormal test samples. Among the successful approaches for anoma ly detection, a distinguished category of methods relies on predicting masked in formation (e.g. patches, future frames, etc.) and leveraging the reconstruction error with respect to the masked information as an abnormality score. Different from related methods, we propose to integrate the reconstruction-based functiona lity into a novel self-supervised predictive architectural building block. The p roposed self-supervised block is generic and can easily be incorporated into var ious state-of-the-art anomaly detection methods. Our block starts with a convolu tional layer with dilated filters, where the center area of the receptive field is masked. The resulting activation maps are passed through a channel attention module. Our block is equipped with a loss that minimizes the reconstruction erro r with respect to the masked area in the receptive field. We demonstrate the gen erality of our block by integrating it into several state-of-the-art frameworks for anomaly detection on image and video, providing empirical evidence that show s considerable performance improvements on MVTec AD, Avenue, and ShanghaiTech. W e release our code as open source at: https://github.com/ristea/sspcab.
****

Safe Self-Refinement for Transformer-Based Domain Adaptation
Tao Sun, Cheng Lu, Tianshuo Zhang, Haibin Ling; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7191-7200
Unsupervised Domain Adaptation (UDA) aims to leverage a label-rich source domain to solve tasks on a related unlabeled target domain. It is a challenging proble m especially when a large domain gap lies between the source and target domains. In this paper we propose a novel solution named SSRT (Safe Self-Refinement for Transformer-based domain adaptation), which brings improvement from two aspects. First, encouraged by the success of vision transformers in various vision tasks , we arm SSRT with a transformer backbone. We find that the combination of visio n transformer with simple adversarial adaptation surpasses best reported Convolu tional Neural Network (CNN)-based results on the challenging DomainNet benchmark , showing its strong transferable feature representation. Second, to reduce the risk of model collapse and improve the effectiveness of knowledge transfer betwe en domains with large gaps, we propose a Safe Self-Refinement strategy. Specific ally, SSRT utilizes predictions of perturbed target domain data to refine the mo del. Since the model capacity of vision transformer is large and predictions in such challenging tasks can be noisy, a safe training mechanism is designed to ad aptively adjust learning configuration. Extensive evaluations are conducted on s everal widely tested UDA benchmarks and SSRT achieves consistently the best perf ormances, including 85.43% on Office-Home, 88.76% on VisDA-2017 and 45.2% on Dom ainNet.
****

Density-Preserving Deep Point Cloud Compression
Yun He, Xinlin Ren, Danhang Tang, Yinda Zhang, Xiangyang Xue, Yanwei Fu; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2333-2342
Local density of point clouds is crucial for representing local details, but has been overlooked by existing point cloud compression methods. To address this, w

e propose a novel deep point cloud compression method that preserves local densi ty information. Our method works in an auto-encoder fashion: the encoder downsam ples the points and learns point-wise features, while the decoder upsamples the points using these features. Specifically, we propose to encode local geometry a nd density with three embeddings: density embedding, local position embedding an d ancestor embedding. During the decoding, we explicitly predict the upsampling factor for each point, and the directions and scales of the upsampled points. To mitigate the clustered points issue in existing methods, we design a novel sub- point convolution layer, and an upsampling block with adaptive scale. Furthermor e, our method can also compress point-wise attributes, such as normal. Extensive qualitative and quantitative results on SemanticKITTI and ShapeNet demonstrate that our method achieves the state-of-the-art rate-distortion trade-off.
****

StyleMesh: Style Transfer for Indoor 3D Scene Reconstructions
Lukas Höllein, Justin Johnson, Matthias Nießner; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6198-6208
We apply style transfer on mesh reconstructions of indoor scenes. This enables V R applications like experiencing 3D environments painted in the style of a favor ite artist. Style transfer typically operates on 2D images, making stylization o f a mesh challenging. When optimized over a variety of poses, stylization patter ns become stretched out and inconsistent in size. On the other hand, model-based 3D style transfer methods exist that allow stylization from a sparse set of ima ges, but they require a network at inference time. To this end, we optimize an e xplicit texture for the reconstructed mesh of a scene and stylize it jointly fro m all available input images. Our depth- and angle-aware optimization leverages surface normal and depth data of the underlying mesh to create a uniform and con sistent stylization for the whole scene. Our experiments show that our method cr eates sharp and detailed results for the complete scene without view-dependent a rtifacts. Through extensive ablation studies, we show that the proposed 3D aware ness enables style transfer to be applied to the 3D domain of a mesh. Our method can be used to render a stylized mesh in real-time with traditional rendering p ipelines.
****

Which Model To Transfer? Finding the Needle in the Growing Haystack
Cedric Renggli, André Susano Pinto, Luka Rimanic, Joan Puigcerver, Carlos Riquel me, Ce Zhang, Mario Lu■i■; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 9205-9214
Transfer learning has been recently popularized as a data-efficient alternative to training models from scratch, in particular for computer vision tasks where i t provides a remarkably solid baseline. The emergence of rich model repositories , such as TensorFlow Hub, enables the practitioners and researchers to unleash t he potential of these models across a wide range of downstream tasks. As these r epositories keep growing exponentially, efficiently selecting a good model for t he task at hand becomes paramount. We provide a formalization of this problem th rough a familiar notion of regret and introduce the predominant strategies, name ly task-agnostic (e.g. ranking models by their ImageNet performance) and task-aw are search strategies (such as linear or kNN evaluation). We conduct a large-sca le empirical study and show that both task-agnostic and task-aware methods can y ield high regret. We then propose a simple and computationally efficient hybrid search strategy which outperforms the existing approaches. We highlight the prac tical benefits of the proposed solution on a set of 19 diverse vision tasks.
****

Fast and Unsupervised Action Boundary Detection for Action Segmentation
Zexing Du, Xue Wang, Guoqing Zhou, Qing Wang; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3323-3332
To deal with the great number of untrimmed videos produced every day, we propose an efficient unsupervised action segmentation method by detecting boundaries, n amed action boundary detection (ABD). In particular, the proposed method has the following advantages: no training stage and low-latency inference. To detect ac tion boundaries, we estimate the similarities across smoothed frames, which inhe

rently have the properties of internal consistency within actions and external d
iscrepancy across actions. Under this circumstance, we successfully transfer the
 boundary detection task into the change point detection based on the similarity
. Then, non-maximum suppression (NMS) is conducted in local windows to select th
e smallest points as candidate boundaries. In addition, a clustering algorithm i
s followed to refine the initial proposals. Moreover, we also extend ABD to the
online setting, which enables real-time action segmentation in long untrimmed vi
deos. By evaluating on four challenging datasets, our method achieves state-of-t
he-art performance. Moreover, thanks to the efficiency of ABD, we achieve the be
st trade-off between the accuracy and the inference time compared with existing
unsupervised approaches.
****

Class-Incremental Learning With Strong Pre-Trained Models
Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasc
oncelos, Rahul Bhotika, Stefano Soatto; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9601-9610
Class-incremental learning (CIL) has been widely studied under the setting of st
arting from a small number of classes (base classes). Instead, we explore an und
erstudied real-world setting of CIL that starts with a strong model pre-trained
on a large number of base classes. We hypothesize that a strong base model can p
rovide a good representation for novel classes and incremental learning can be d
one with small adaptations. We propose a 2-stage training scheme, i) feature aug
mentation - cloning part of the backbone and fine-tuning it on the novel data, a
nd ii) fusion - combining the base and novel classifiers into a unified classifi
er. Experiments show that the proposed method significantly outperforms state-of
-the-art CIL methods on the large-scale ImageNet dataset (e.g. +10% overall accu
racy than the best). We also propose and analyze understudied practical CIL scen
arios, such as base-novel overlap with distribution shift. Our proposed method i
s robust and generalizes to all analyzed CIL settings.
****

Robust Optimization As Data Augmentation for Large-Scale Graphs
Kezhi Kong, Guohao Li, Mucong Ding, Zuxuan Wu, Chen Zhu, Bernard Ghanem, Gavin T
aylor, Tom Goldstein; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2022, pp. 60-69
Data augmentation helps neural networks generalize better by enlarging the train
ing set, but it remains an open question how to effectively augment graph data t
o enhance the performance of GNNs (Graph Neural Networks). While most existing g
raph regularizers focus on manipulating graph topological structures by adding/r
emoving edges, we offer a method to augment node features for better performance
. We propose FLAG (Free Large-scale Adversarial Augmentation on Graphs), which i
teratively augments node features with gradient-based adversarial perturbations
during training. By making the model invariant to small fluctuations in input da
ta, our method helps models generalize to out-of-distribution samples and boosts
 model performance at test time. FLAG is a general-purpose approach for graph da
ta, which universally works in node classification, link prediction, and graph c
lassification tasks. FLAG is also highly flexible and scalable, and is deployabl
e with arbitrary GNN backbones and large-scale datasets. We demonstrate the effi
cacy and stability of our method through extensive experiments and ablation stud
ies. We also provide intuitive observations for a deeper understanding of our me
thod. We open source our implementation at https://github.com/devnkong/FLAG.
****

Robust Structured Declarative Classifiers for 3D Point Clouds: Defending Adversa
rial Attacks With Implicit Gradients
Kaidong Li, Ziming Zhang, Cuncong Zhong, Guanghui Wang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1529
4-15304
Deep neural networks for 3D point cloud classification, such as PointNet, have b
een demonstrated to be vulnerable to adversarial attacks. Current adversarial de
fenders often learn to denoise the (attacked) point clouds by reconstruction, an
d then feed them to the classifiers as input. In contrast to the literature, we

propose a family of robust structured declarative classifiers for point cloud classification, where the internal constrained optimization mechanism can effectively defend adversarial attacks through implicit gradients. Such classifiers can be formulated using a bilevel optimization framework. We further propose an effective and efficient instantiation of our approach, namely, Lattice Point Classifier (LPC), based on structured sparse coding in the permutohedral lattice and 2D convolutional neural networks (CNNs) that is end-to-end trainable. We demonstrate state-of-the-art robust point cloud classification performance on ModelNet40 and ScanNet under seven different attackers. For instance, we achieve 89.51% and 83.16% test accuracy on each dataset under the recent JGBA attacker that outperforms DUP-Net and IF-Defense with PointNet by 70%. Demo code is available at https://zhang-vislab.github.io.

****

PhotoScene: Photorealistic Material and Lighting Transfer for Indoor Scenes

Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18562-18571

Most indoor 3D scene reconstruction methods focus on recovering 3D geometry and scene layout. In this work, we go beyond this to propose PhotoScene, a framework that takes input image(s) of a scene along with approximately aligned CAD geometry (either reconstructed automatically or manually specified) and builds a photorealistic digital twin with high-quality materials and similar lighting. We model scene materials using procedural material graphs; such graphs represent photorealistic and resolution-independent materials. We optimize the parameters of these graphs and their texture scale and rotation, as well as the scene lighting to best match the input image via a differentiable rendering layer. We evaluate our technique on objects and layout reconstructions from ScanNet, SUN RGB-D and stock photographs, and demonstrate that our method reconstructs high-quality, fully relightable 3D scenes that can be re-rendered under arbitrary viewpoints, zooms and lighting.

****

Improving the Transferability of Targeted Adversarial Examples Through Object-Based Diverse Input

Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, Changick Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15244-15253

The transferability of adversarial examples allows the deception on black-box models, and transfer-based targeted attacks have attracted a lot of interest due to their practical applicability. To maximize the transfer success rate, adversarial examples should avoid overfitting to the source model, and image augmentation is one of the primary approaches for this. However, prior works utilize simple image transformations such as resizing, which limits input diversity. To tackle this limitation, we propose the object-based diverse input (ODI) method that draws an adversarial image on a 3D object and induces the rendered image to be classified as the target class. Our motivation comes from the humans' superior perception of an image printed on a 3D object. If the image is clear enough, humans can recognize the image content in a variety of viewing conditions. Likewise, if an adversarial example looks like the target class to the model, the model should also classify the rendered image of the 3D object as the target class. The ODI method effectively diversifies the input by leveraging an ensemble of multiple source objects and randomizing viewing conditions. In our experimental results on the ImageNet-Compatible dataset, this method boosts the average targeted attack success rate from 28.3% to 47.0% compared to the state-of-the-art methods. We also demonstrate the applicability of the ODI method to adversarial examples on the face verification task and its superior performance improvement. Our code is available at https://github.com/dreamflake/ODI.

****

IRON: Inverse Rendering by Optimizing Neural SDFs and Materials From Photometric Images

Kai Zhang, Fujun Luan, Zhengqi Li, Noah Snavely; Proceedings of the IEEE/CVF Con

ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5565-5574
We propose a neural inverse rendering pipeline called IRON that operates on photometric images and outputs high-quality 3D content in the format of triangle meshes and material textures readily deployable in existing graphics pipelines. We propose a neural inverse rendering pipeline called IRON that operates on photometric images and outputs high-quality 3D content in the format of triangle meshes and material textures readily deployable in existing graphics pipelines. Our method adopts neural representations for geometry as signed distance fields (SDFs) and materials during optimization to enjoy their flexibility and compactness, and features a hybrid optimization scheme for neural SDFs: first, optimize using a volumetric radiance field approach to recover correct topology, then optimize further using edge-aware physics-based surface rendering for geometry refinement and disentanglement of materials and lighting. In the second stage, we also draw inspiration from mesh-based differentiable rendering, and design a novel edge sampling algorithm for neural SDFs to further improve performance. We show that our IRON achieves significantly better inverse rendering quality compared to prior works.
****

ObjectFolder 2.0: A Multisensory Object Dataset for Sim2Real Transfer
Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10598-10608
Objects play a crucial role in our everyday activities. Though multisensory object-centric learning has shown great potential lately, the modeling of objects in prior work is rather unrealistic. ObjectFolder 1.0 is a recent dataset that introduces 100 virtualized objects with visual, auditory, and tactile sensory data. However, the dataset is small in scale and the multisensory data is of limited quality, hampering generalization to real-world scenarios. We present ObjectFolder 2.0, a large-scale, multisensory dataset of common household objects in the form of implicit neural representations that significantly enhances ObjectFolder 1.0 in three aspects. First, our dataset is 10 times larger in the amount of objects and orders of magnitude faster in rendering time. Second, we significantly improve the multisensory rendering quality for all three modalities. Third, we show that models learned from virtual objects in our dataset successfully transfer to their real-world counterparts in three challenging tasks: object scale estimation, contact localization, and shape reconstruction. ObjectFolder 2.0 offers a new path and testbed for multisensory learning in computer vision and robotics. The dataset is available at https://github.com/rhgao/ObjectFolder.
****

Versatile Multi-Modal Pre-Training for Human-Centric Perception
Fangzhou Hong, Liang Pan, Zhongang Cai, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16156-16166
Human-centric perception plays a vital role in vision and graphics. But their data annotations are prohibitively expensive. Therefore, it is desirable to have a versatile pre-train model that serves as a foundation for data-efficient downstream tasks transfer. To this end, we propose the Human-Centric Multi-Modal Contrastive Learning framework HCMoCo that leverages the multi-modal nature of human data (e.g. RGB, depth, 2D keypoints) for effective representation learning. The objective comes with two main challenges: dense pre-train for multi-modality data, efficient usage of sparse human priors. To tackle the challenges, we design the novel Dense Intra-sample Contrastive Learning and Sparse Structure-aware Contrastive Learning targets by hierarchically learning a modal-invariant latent space featured with continuous and ordinal feature distribution and structure-aware semantic consistency. HCMoCo provides pre-train for different modalities by combining heterogeneous datasets, which allows efficient usage of existing task-specific human data. Extensive experiments on four downstream tasks of different modalities demonstrate the effectiveness of HCMoCo, especially under data-efficient settings (7.16% and 12% improvement on DensePose Estimation and Human Parsing). Moreover, we demonstrate the versatility of HCMoCo by exploring cross-modality

supervision and missing-modality inference, validating its strong ability in cross-modal association and reasoning.
****

360MonoDepth: High-Resolution 360deg Monocular Depth Estimation
Manuel Rey-Area, Mingze Yuan, Christian Richardt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3762-3772
360deg cameras can capture complete environments in a single shot, which makes 360deg imagery alluring in many computer vision tasks. However, monocular depth estimation remains a challenge for 360deg data, particularly for high resolutions like 2K (2048x1024) and beyond that are important for novel-view synthesis and virtual reality applications. Current CNN-based methods do not support such high resolutions due to limited GPU memory. In this work, we propose a flexible framework for monocular depth estimation from high-resolution 360deg images using tangent images. We project the 360deg input image onto a set of tangent planes that produce perspective views, which are suitable for the latest, most accurate state-of-the-art perspective monocular depth estimators. To achieve globally consistent disparity estimates, we recombine the individual depth estimates using deformable multi-scale alignment followed by gradient-domain blending. The result is a dense, high-resolution 360deg depth map with a high level of detail, also for outdoor scenes which are not supported by existing methods. Our source code and data are available at https://manurare.github.io/360monodepth/.
****

Splicing ViT Features for Semantic Appearance Transfer
Narek Tumanyan, Omer Bar-Tal, Shai Bagon, Tali Dekel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10748-10757
We present a method for semantically transferring the visual appearance of one natural image to another. Specifically, our goal is to generate an image in which objects in a source structure image are "painted" with the visual appearance of their semantically related objects in a target appearance image. Our method works by training a generator given only a single structure/appearance image pair as input. To integrate semantic information into our framework---a pivotal component in tackling this task---our key idea is to leverage a pre-trained and fixed Vision Transformer (ViT) model which serves as an external semantic prior. Specifically, we derive novel representations of structure and appearance extracted from deep ViT features, untwisting them from the learned self-attention modules. We then establish an objective function that splices the desired structure and appearance representations, interweaving them together in the space of ViT features. Our framework, which we term "Splice", does not involve adversarial training, nor does it require any additional input information such as semantic segmentation or correspondences, and can generate high resolution results, e.g., work in HD. We demonstrate high quality results on a variety of in-the-wild image pairs, under significant variations in the number of objects, their pose and appearance.
****

Contrastive Regression for Domain Adaptation on Gaze Estimation
Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, Teng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19376-19385
Appearance-based Gaze Estimation leverages deep neural networks to regress the gaze direction from monocular images and achieve impressive performance. However, its success depends on expensive and cumbersome annotation capture. When lacking precise annotation, the large domain gap hinders the performance of trained models on new domains. In this paper, we propose a novel gaze adaptation approach, namely Contrastive Regression Gaze Adaptation (CRGA), for generalizing gaze estimation on the target domain in an unsupervised manner. CRGA leverages the Contrastive Domain Generalization (CDG) module to learn the stable representation from the source domain and leverages the Contrastive Self-training Adaptation (CSA) module to learn from the pseudo labels on the target domain. The core of both CDG and CSA is the Contrastive Regression (CR) loss, a novel contrastive loss for

regression by pulling features with closer gaze directions closer together while pushing features with farther gaze directions farther apart. Experimentally, we choose ETH-XGAZE and Gaze-360 as the source domain and test the domain generalization and adaptation performance on MPIIGAZE, RT-GENE, GazeCapture, EyeDiap respectively. The results demonstrate that our CRGA achieves remarkable performance improvement compared with the baseline models and also outperforms the state-of-the-art domain adaptation approaches on gaze adaptation tasks.

****

MUSE-VAE: Multi-Scale VAE for Environment-Aware Long Term Trajectory Prediction

Mihee Lee, Samuel S. Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, Vladimir Pavlovic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2221-2230

Accurate long-term trajectory prediction in complex scenes, where multiple agents (e.g., pedestrians or vehicles) interact with each other and the environment while attempting to accomplish diverse and often unknown goals, is a challenging stochastic forecasting problem. In this work, we propose MUSE-VAE, a new probabilistic modeling framework based on a cascade of Conditional VAEs, which tackles the long-term, uncertain trajectory prediction task using a coarse-to-fine multi-factor forecasting architecture. In its Macro stage, the model learns a joint pixel-space representation of two key factors, the underlying environment and the agent movements, to predict the long and short term motion goals. Conditioned on them, the Micro stage learns a fine-grained spatio-temporal representation for the prediction of individual agent trajectories. The VAE backbones across the two stages make it possible to naturally account for the joint uncertainty at both levels of granularity. As a result, MUSE-VAE offers diverse and simultaneously more accurate predictions compared to the current state-of-the-art. We demonstrate these assertions through a comprehensive set of experiments on nuScenes and SDD benchmarks as well as PFSD, a new synthetic dataset, which challenges the forecasting ability of models on complex agent-environment interaction scenarios.

****

Multi-View Consistent Generative Adversarial Networks for 3D-Aware Image Synthesis

Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18450-18459

3D-aware image synthesis aims to generate images of objects from multiple views by learning a 3D representation. However, one key challenge remains: existing approaches lack geometry constraints, hence usually fail to generate multi-view consistent images. To address this challenge, we propose Multi-View Consistent Generative Adversarial Networks (MVCGAN) for high-quality 3D-aware image synthesis with geometry constraints. By leveraging the underlying 3D geometry information of generated images, i.e., depth and camera transformation matrix, we explicitly establish stereo correspondence between views to perform multi-view joint optimization. In particular, we enforce the photometric consistency between pairs of views and integrate a stereo mixup mechanism into the training process, encouraging the model to reason about the correct 3D shape. Besides, we design a two-stage training strategy with feature-level multi-view joint optimization to improve the image quality. Extensive experiments on three datasets demonstrate that MVCGAN achieves the state-of-the-art performance for 3D-aware image synthesis.

****

Putting People in Their Place: Monocular Regression of 3D People in Depth

Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13243-13252

Given an image with multiple people, our goal is to directly regress the pose and shape of all the people as well as their relative depth. Inferring the depth of a person in an image, however, is fundamentally ambiguous without knowing their height. This is particularly problematic when the scene contains people of very different sizes, e.g. from infants to adults. To solve this, we need several things. First, we develop a novel method to infer the poses and depth of multiple

people in a single image. While previous work that estimates multiple people does so by reasoning in the image plane, our method, called BEV, adds an additional imaginary Bird's-Eye-View representation to explicitly reason about depth. BEV reasons simultaneously about body centers in the image and in depth and, by combing these, estimates 3D body position. Unlike prior work, BEV is a single-shot method that is end-to-end differentiable. Second, height varies with age, making it impossible to resolve depth without also estimating the age of people in the image. To do so, we exploit a 3D body model space that lets BEV infer shapes from infants to adults. Third, to train BEV, we need a new dataset. Specifically, we create a "Relative Human" (RH) dataset that includes age labels and relative depth relationships between the people in the images. Extensive experiments on RH and AGORA demonstrate the effectiveness of the model and training scheme. BEV outperforms existing methods on depth reasoning, child shape estimation, and robustness to occlusion. The code and dataset are released for research purposes.

****

POCO: Point Convolution for Surface Reconstruction

Alexandre Boulch, Renaud Marlet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6302-6314

Implicit neural networks have been successfully used for surface reconstruction from point clouds. However, many of them face scalability issues as they encode the isosurface function of a whole object or scene into a single latent vector. To overcome this limitation, a few approaches infer latent vectors on a coarse regular 3D grid or on 3D patches, and interpolate them to answer occupancy queries. In doing so, they loose the direct connection with the input points sampled on the surface of objects, and they attach information uniformly in space rather than where it matters the most, i.e., near the surface. Besides, relying on fixed patch sizes may require discretization tuning. To address these issues, we propose to use point cloud convolutions and compute latent vectors at each input point. We then perform a learning-based interpolation on nearest neighbors using inferred weights. Experiments on both object and scene datasets show that our approach significantly outperforms other methods on most classical metrics, producing finer details and better reconstructing thinner volumes. The code is available at https://github.com/valeoai/POCO

****

Memory-Augmented Non-Local Attention for Video Super-Resolution

Jiyang Yu, Jingen Liu, Liefeng Bo, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17834-17843

In this paper, we propose a simple yet effective video super-resolution method that aims at generating high-fidelity high-resolution (HR) videos from low-resolution (LR) ones. Previous methods predominantly leverage temporal neighbor frames to assist the super-resolution of the current frame. Those methods achieve limited performance as they suffer from the challenges in spatial frame alignment and the lack of useful information from similar LR neighbor frames. In contrast, we devise a cross-frame non-local attention mechanism that allows video super-resolution without frame alignment, leading to being more robust to large motions in the video. In addition, to acquire general video prior information beyond neighbor frames, and to compensate for the information loss caused by large motions, we design a novel memory-augmented attention module to memorize general video details during the super-resolution training. We have thoroughly evaluated our work on various challenging datasets. Compared to other recent video super-resolution approaches, our method not only achieves significant performance gains on large motion videos but also shows better generalization. Our source code and the new Parkour benchmark dataset will be released.

****

Neural Texture Extraction and Distribution for Controllable Person Image Synthesis

Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, Thomas H. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13535-13544

We deal with the controllable person image synthesis task which aims to re-rende

r a human from a reference image with explicit control over body pose and appear
ance. Observing that person images are highly structured, we propose to generate
 desired images by extracting and distributing semantic entities of reference im
ages. To achieve this goal, a neural texture extraction and distribution operati
on based on double attention is described. This operation first extracts semanti
c neural textures from reference feature maps. Then, it distributes the extracte
d neural textures according to the spatial distributions learned from target pos
es. Our model is trained to predict human images in arbitrary poses, which encou
rages it to extract disentangled and expressive neural textures representing the
 appearance of different semantic entities. The disentangled representation furt
her enables explicit appearance control. Neural textures of different reference
 images can be fused to control the appearance of the interested areas. Experimen
tal comparisons show the superiority of the proposed model. Code is available at
 https://github.com/RenYurui/Neural-Texture-Extraction-Distribution.
****

Classification-Then-Grounding: Reformulating Video Scene Graphs As Temporal Bipa
rtite Graphs

Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, Jun Xiao; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1949
7-19506

Today's VidSGG models are all proposal-based methods, i.e., they first generate
 numerous paired subject-object snippets as proposals, and then conduct predicate
 classification for each proposal. In this paper, we argue that this prevalent p
roposal-based framework has three inherent drawbacks: 1) The ground-truth predic
ate labels for proposals are partially correct. 2) They break the high-order rel
ations among different predicate instances of a same subject-object pair. 3) Vid
SGG performance is upper-bounded by the quality of the proposals. To this end, w
e propose a new classification-then-grounding framework for VidSGG, which can av
oid all the three overlooked drawbacks. Meanwhile, under this framework, we refo
rmulate the video scene graphs as temporal bipartite graphs, where the entities
 and predicates are two types of nodes with time slots, and the edges denote diff
erent semantic roles between these nodes. This formulation takes full advantage
 of our new framework. Accordingly, we further propose a novel BIpartite Graph ba
sed SGG model: BIG. It consists of a classification stage and a grounding stage,
 where the former aims to classify the categories of all the nodes and the edges
, and the latter tries to localize the temporal location of each relation instan
ce. Extensive ablations on two VidSGG datasets have attested to the effectivenes
s of our framework and BIG. Code is available at https://github.com/Dawn-LX/VidS
GG-BIG.
****

Transformer-Empowered Multi-Scale Contextual Matching and Aggregation for Multi-
Contrast MRI Super-Resolution

Guangyuan Li, Jun Lv, Yapeng Tian, Qi Dou, Chengyan Wang, Chenliang Xu, Jing Qin
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 20636-20645

Magnetic resonance imaging (MRI) can present multi-contrast images of the same a
natomical structures, enabling multi-contrast super-resolution (SR) techniques.
Compared with SR reconstruction using a single-contrast, multi-contrast SR recon
struction is promising to yield SR images with higher quality by leveraging dive
rse yet complementary information embedded in different imaging modalities. Howe
ver, existing methods still have two shortcomings: (1) they neglect that the mul
ti-contrast features at different scales contain different anatomical details an
d hence lack effective mechanisms to match and fuse these features for better re
construction; and (2) they are still deficient in capturing long-range dependenc
ies, which are essential for the regions with complicated anatomical structures.
 We propose a novel network to comprehensively address these problems by develop
ing a set of innovative Transformer-empowered multi-scale contextual matching an
d aggregation techniques; we call it McMRSR. Firstly, we tame transformers to mo
del long-range dependencies in both reference and target images. Then, a new mul
ti-scale contextual matching method is proposed to capture corresponding context

s from reference features at different scales. Furthermore, we introduce a multi-scale aggregation mechanism to gradually and interactively aggregate multi-scale matched features for reconstructing the target SR MR image. Extensive experiments demonstrate that our network outperforms state-of-the-art approaches and has great potential to be applied in clinical practice.

****

GazeOnce: Real-Time Multi-Person Gaze Estimation

Mingfang Zhang, Yunfei Liu, Feng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4197-4206

Appearance-based gaze estimation aims to predict the 3D eye gaze direction from a single image. While recent deep learning-based approaches have demonstrated excellent performance, they usually assume one calibrated face in each input image and cannot output multi-person gaze in real time. However, simultaneous gaze estimation for multiple people in the wild is necessary for real-world applications. In this paper, we propose the first one-stage end-to-end gaze estimation method, GazeOnce, which is capable of simultaneously predicting gaze directions for multiple faces (>10) in an image. In addition, we design a sophisticated data generation pipeline and propose a new dataset, MPSGaze, which contains full images of multiple people with 3D gaze ground truth. Experimental results demonstrate that our unified framework not only offers a faster speed, but also provides a lower gaze estimation error compared with state-of-the-art methods. This technique can be useful in real-time applications with multiple users.

****

GateHUB: Gated History Unit With Background Suppression for Online Action Detection

Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, Mei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19925-19934

Online action detection is the task of predicting the action as soon as it happens in a streaming video. A major challenge is that the model does not have access to the future and has to solely rely on the history, i.e., the frames observed so far, to make predictions. It is therefore important to accentuate parts of the history that are more informative to the prediction of the current frame. We present GateHUB, Gated History Unit with Background Suppression, that comprises a novel position-guided gated cross-attention mechanism to enhance or suppress parts of the history as per how informative they are for current frame prediction. GateHUB further proposes Future-augmented History (FaH) to make history features more informative by using subsequently observed frames when available. In a single unified framework, GateHUB integrates the transformer's ability of long-range temporal modeling and the recurrent model's capacity to selectively encode relevant information. GateHUB also introduces a background suppression objective to further mitigate false positive background frames that closely resemble the action frames. Extensive validation on three benchmark datasets, THUMOS, TVSeries, and HDD, demonstrates that GateHUB significantly outperforms all existing methods and is also more efficient than the existing best work. Furthermore, a flow-free version of GateHUB is able to achieve higher or close accuracy at 2.8x higher frame rate compared to all existing methods that require both RGB and optical flow information for prediction.

****

Few-Shot Font Generation by Learning Fine-Grained Local Styles

Licheng Tang, Yiyang Cai, Jiaming Liu, Zhibin Hong, Mingming Gong, Minhu Fan, Junyu Han, Jingtuo Liu, Errui Ding, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7895-7904

Few-shot font generation (FFG), which aims to generate a new font with a few examples, is gaining increasing attention due to the significant reduction in labor cost. A typical FFG pipeline considers characters in a standard font library as content glyphs and transfers them to a new target font by extracting style information from the reference glyphs. Most existing solutions explicitly disentangle content and style of reference glyphs globally or component-wisely. However, the style of glyphs mainly lies in the local details, i.e. the styles of radicals

, components, and strokes together depict the style of a glyph. Therefore, even a single character can contain different styles distributed over spatial locations. In this paper, we propose a new font generation approach by learning 1) the fine-grained local styles from references, and 2) the spatial correspondence between the content and reference glyphs. Therefore each spatial location in the content glyph can be assigned with the right fine-grained style. To this end, we adopt cross-attention over the representation of the content glyphs as the queries and the representations of the reference glyphs as the keys and values. Instead of explicitly disentangling global or component-wise modeling, the cross attention mechanism can attend to the right local styles in the reference glyphs and aggregates the reference styles into a fine-grained style representation for the given content glyphs. The experiments show that the proposed method outperforms the state-of-the-art methods in FFG. In particular, the user studies also demonstrate the style consistency of our approach is significantly outperforms previous methods.

****

Bridging Video-Text Retrieval With Multiple Choice Questions

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16167-16176

Pre-training a model to learn transferable video-text representation for retrieval has attracted a lot of attention in recent years. Previous dominant works mainly adopt two separate encoders for efficient retrieval, but ignore local associations between videos and texts. Another line of research uses a joint encoder to interact video with texts, but results in low efficiency since each text-video pair needs to be fed into the model. In this work, we enable fine-grained video-text interactions while maintaining high efficiency for retrieval via a novel pretext task, dubbed as Multiple Choice Questions (MCQ), where a parametric module BridgeFormer is trained to answer the "questions" constructed by the text features via resorting to the video features. Specifically, we exploit the rich semantics of text (i.e., nouns and verbs) to build questions, with which the video encoder can be trained to capture more regional content and temporal dynamics. In the form of questions and answers, the semantic associations between local video-text features can be properly established. BridgeFormer is able to be removed for downstream retrieval, rendering an efficient and flexible model with only two encoders. Our method outperforms state-of-the-art methods on the popular text-to-video retrieval task in five datasets with different experimental setups (i.e., zero-shot and fine-tune), including HowTo100M (one million videos). We further conduct zero-shot action recognition, which can be cast as video-to-text retrieval, and our approach also significantly surpasses its counterparts. As an additional benefit, our method achieves competitive results with much shorter pre-training videos on single-modality downstream tasks, e.g., action recognition with linear evaluation.

****

Depth-Aware Generative Adversarial Network for Talking Head Video Generation

Fa-Ting Hong, Longhao Zhang, Li Shen, Dan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3397-3406

Talking head video generation aims to produce a synthetic human face video that contains the identity and pose information respectively from a given source image and a driving video. Existing works for this task heavily rely on 2D representations (e.g. appearance and motion) learned from the input images. However, dense 3D facial geometry (e.g. pixel-wise depth) is extremely important for this task as it is particularly beneficial for us to essentially generate accurate 3D face structures and distinguish noisy information from the possibly cluttered background. Nevertheless, dense 3D geometry annotations are prohibitively costly for videos and are typically not available for this video generation task. In this paper, we introduce a self-supervised face-depth learning method to automatically recover dense 3D facial geometry (i.e. depth) from the face videos without the requirement of any expensive 3D annotation data. Based on the learned dense depth maps, we further propose to leverage them to estimate sparse facial keypoints

that capture the critical movement of the human head. In a more dense way, the depth is also utilized to learn 3D-aware cross-modal (i.e. appearance and depth) attention to guide the generation of motion fields for warping source image representations. All these contributions compose a novel depth-aware generative adversarial network (DaGAN) for talking head generation. Extensive experiments conducted demonstrate that our proposed method can generate highly realistic faces, and achieve significant results on the unseen human faces.

****

Dual-Path Image Inpainting With Auxiliary GAN Inversion

Wentao Wang, Li Niu, Jianfu Zhang, Xue Yang, Liqing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11421-11430

Deep image inpainting can inpaint a corrupted image using a feed-forward inference, but still fails to handle large missing area or complex semantics. Recently, GAN inversion based inpainting methods propose to leverage semantic information in pretrained generator (e.g., StyleGAN) to solve the above issues. Different from feed-forward methods, they seek for a closest latent code to the corrupted image and feed it to a pretrained generator. However, inferring the latent code is either time-consuming or inaccurate. In this paper, we develop a dual-path inpainting network with inversion path and feed-forward path, in which inversion path provides auxiliary information to help feed-forward path. We also design a novel deformable fusion module to align the feature maps in two paths. Experiments on FFHQ and LSUN demonstrate that our method is effective in solving the aforementioned problems while producing more realistic results than state-of-the-art methods.

****

DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16515-16525

Synthesizing high-quality realistic images from text descriptions is a challenging task. Existing text-to-image Generative Adversarial Networks generally employ a stacked architecture as the backbone yet still remain three flaws. First, the stacked architecture introduces the entanglements between generators of different image scales. Second, existing studies prefer to apply and fix extra networks in adversarial learning for text-image semantic consistency, which limits the supervision capability of these networks. Third, the cross-modal attention-based text-image fusion that widely adopted by previous works is limited on several special image scales because of the computational cost. To these ends, we propose a simpler but more effective Deep Fusion Generative Adversarial Networks (DF-GAN). To be specific, we propose: (i) a novel one-stage text-to-image backbone that directly synthesizes high-resolution images without entanglements between different generators, (ii) a novel Target-Aware Discriminator composed of Matching-Aware Gradient Penalty and One-Way Output, which enhances the text-image semantic consistency without introducing extra networks, (iii) a novel deep text-image fusion block, which deepens the fusion process to make a full fusion between text and visual features. Compared with current state-of-the-art methods, our proposed DF-GAN is simpler but more efficient to synthesize realistic and text-matching images and achieves better performance on widely used datasets. Code is available at https://github.com/tobran/DF-GAN.

****

Generative Flows With Invertible Attentions

Rhea Sanjay Sukthanker, Zhiwu Huang, Suryansh Kumar, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11234-11243

Flow-based generative models have shown an excellent ability to explicitly learn the probability density function of data via a sequence of invertible transformations. Yet, learning attentions in generative flows remains understudied, while it has made breakthroughs in other domains. To fill the gap, this paper introduces two types of invertible attention mechanisms, i.e., map-based and transforme

r-based attentions, for both unconditional and conditional generative flows. The key idea is to exploit a masked scheme of these two attentions to learn long-range data dependencies in the context of generative flows. The masked scheme allows for invertible attention modules with tractable Jacobian determinants, enabling its seamless integration at any positions of the flow-based models. The proposed attention mechanisms lead to more efficient generative flows, due to their capability of modeling the long-term data dependencies. Evaluation on multiple image synthesis tasks shows that the proposed attention flows result in efficient models and compare favorably against the state-of-the-art unconditional and conditional generative flows.
****

Clipped Hyperbolic Classifiers Are Super-Hyperbolic Classifiers
Yunhui Guo, Xudong Wang, Yubei Chen, Stella X. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11-20
Hyperbolic space can naturally embed hierarchies, unlike Euclidean space. Hyperbolic Neural Networks (HNNs) exploit such representational power by lifting Euclidean features into hyperbolic space for classification, outperforming Euclidean neural networks (ENNs) on datasets with known semantic hierarchies. However, HNNs underperform ENNs on standard benchmarks without clear hierarchies, greatly restricting HNNs' applicability in practice. Our key insight is that HNNs' poorer general classification performance results from vanishing gradients during backpropagation, caused by their hybrid architecture connecting Euclidean features to a hyperbolic classifier. We propose an effective solution by simply clipping the Euclidean feature magnitude while training HNNs. Our experiments demonstrate that clipped HNNs become super-hyperbolic classifiers: They are not only consistently better than HNNs which already outperform ENNs on hierarchical data, but also on-par with ENNs on MNIST, CIFAR10, CIFAR100 and ImageNet benchmarks, with better adversarial robustness and out-of-distribution detection.
****

Estimating Fine-Grained Noise Model via Contrastive Learning
Yunhao Zou, Ying Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12682-12691
Image denoising has achieved unprecedented progress as great efforts have been made to exploit effective deep denoisers. To improve the denoising performance in real-world, two typical solutions are used in recent trends: devising better noise models for the synthesis of more realistic training data, and estimating noise level function to guide non-blind denoisers. In this work, we combine both noise modeling and estimation, and propose an innovative noise model estimation and noise synthesis pipeline for realistic noisy image generation. Specifically, our model learns a noise estimation model with fine-grained statistical noise model in a contrastive manner. Then, we use the estimated noise parameters to model camera-specific noise distribution, and synthesize realistic noisy training data. The most striking thing for our work is that by calibrating noise models of several sensors, our model can be extended to predict other cameras. In other words, we can estimate camera-specific noise models for unknown sensors with only testing images, without any laborious calibration frames or paired noisy/clean data. The proposed pipeline endows deep denoisers with competitive performances with state-of-the-art real noise modeling methods.
****

DiffPoseNet: Direct Differentiable Camera Pose Estimation
Chethan M. Parameshwara, Gokul Hari, Cornelia Fermüller, Nitin J. Sanket, Yiannis Aloimonos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6845-6854
Current deep neural network approaches for camera pose estimation rely on scene structure for 3D motion estimation, but this decreases the robustness and thereby makes cross-dataset generalization difficult. In contrast, classical approaches to structure from motion estimate 3D motion utilizing optical flow and then compute depth. Their accuracy, however, depends strongly on the quality of the optical flow. To avoid this issue, direct methods have been proposed, which separate 3D motion from depth estimation but compute 3D motion using only image gradien

ts in the form of normal flow. In this paper, we introduce a network NFlowNet, f
or normal flow estimation which is used to enforce robust and direct constraints
. In particular, normal flow is used to estimate relative camera pose based on t
he cheirality (depth positivity) constraint. We achieve this by formulating the
optimization problem as a differentiable cheirality layer, which allows for end-
to-end learning of camera pose. We perform extensive qualitative and quantitativ
e evaluation of the proposed DiffPoseNet's sensitivity to noise and its generali
zation across datasets. We compare our approach to existing state-of-the-art met
hods on KITTI, TartanAir, and TUM-RGBD datasets
****

The Flag Median and FlagIRLS
Nathan Mankovich, Emily J. King, Chris Peterson, Michael Kirby; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
pp. 10339-10347
Finding prototypes (e.g., mean and median) for a dataset is central to a number
of common machine learning algorithms. Subspaces have been shown to provide usef
ul, robust representations for datasets of images, videos and more. Since subspa
ces correspond to points on a Grassmann manifold, one is led to consider the ide
a of a subspace prototype for a Grassmann-valued dataset. While a number of diff
erent subspace prototypes have been described, the calculation of some of these
prototypes has proven to be computationally expensive while other prototypes are
 affected by outliers and produce highly imperfect clustering on noisy data. Thi
s work proposes a new subspace prototype, the flag median, and introduces the Fl
agIRLS algorithm for its calculation. We provide evidence that the flag median i
s robust to outliers and can be used effectively in algorithms like Linde-Buzo-G
rey (LBG) to produce improved clusterings on Grassmannians. Numerical experiment
s include a synthetic dataset, the MNIST handwritten digits dataset, the Mind's
Eye video dataset and the UCF YouTube action dataset. The flag median is compare
d the other leading algorithms for computing prototypes on the Grassmannian, nam
ely, the l_2-median and to the flag mean. We find that using FlagIRLS to compute
 the flag median converges in 4 iterations on a synthetic dataset. We also see t
hat Grassmannian LBG with a codebook size of 20 and using the flag median produc
es at least a 10% improvement in cluster purity over Grassmannian LBG using the
flag mean or l_2-median on the Mind's Eye dataset.
****

Implicit Feature Decoupling With Depthwise Quantization
Iordanis Fostiropoulos, Barry Boehm; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2022, pp. 396-405
Quantization has been applied to multiple domains in Deep Neural Networks (DNNs)
. We propose Depthwise Quantization (DQ) where quantization is applied to a deco
mposed sub-tensor along the feature axis of weak statistical dependence. The fea
ture decomposition leads to an exponential increase in representation capacity w
ith a linear increase in memory and parameter cost. In addition, DQ can be direc
tly applied to existing encoder-decoder frameworks without modification of the D
NN architecture. We use DQ in the context of Hierarchical Auto-Encoders and trai
n end-to-end on an image feature representation. We provide an analysis of the c
ross-correlation between spatial and channel features and propose a decompositio
n of the image feature representation along the channel axis. The improved perfo
rmance of the depthwise operator is due to the increased representation capacity
 from implicit feature decoupling. We evaluate DQ on the likelihood estimation t
ask, where it outperforms the previous state-of-the-art on CIFAR-10, ImageNet-32
 and ImageNet-64. We progressively train with increasing image size a single hie
rarchical model that uses 69% fewer parameters and has faster convergence than t
he previous work.
****

Graph-Context Attention Networks for Size-Varied Deep Graph Matching
Zheheng Jiang, Hossein Rahmani, Plamen Angelov, Sue Black, Bryan M. Williams; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2022, pp. 2343-2352
Deep learning for graph matching has received growing interest and developed rap

idly in the past decade. Although recent deep graph matching methods have shown excellent performance on matching between graphs of equal size in the computer vision area, the size-varied graph matching problem, where the number of keypoints in the images of the same category may vary due to occlusion, is still an open and challenging problem. To tackle this, we firstly propose to formulate the combinatorial problem of graph matching as an Integer Linear Programming (ILP) problem, which is more flexible and efficient to facilitate comparing graphs of varied sizes. A novel Graph-context Attention Network (GCAN), which jointly capture intrinsic graph structure and cross-graph information for improving the discrimination of node features, is then proposed and trained to resolve this ILP problem with node correspondence supervision. We further show that the proposed GCAN model is efficient to resolve the graph-level matching problem and is able to automatically learn node-to-node similarity via graph-level matching. The proposed approach is evaluated on three public keypoint-matching datasets and one graph-matching dataset for blood vessel patterns, with experimental results showing its superior performance over existing state-of-the-art algorithms on the keypoint and graph-level matching tasks.

****

FENeRF: Face Editing in Neural Radiance Fields

Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, Jue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7672-7682

Previous portrait image generation methods roughly fall into two categories: 2D GANs and 3D-aware GANs. 2D GANs can generate high fidelity portraits but with low view consistency. 3D-aware GAN methods can maintain view consistency but their generated images are not locally editable. To overcome these limitations, we propose FENeRF, a 3D-aware generator that can produce view-consistent and locally-editable portrait images. Our method uses two decoupled latent codes to generate corresponding facial semantics and texture in a spatial-aligned 3D volume with shared geometry. Benefiting from such underlying 3D representation, FENeRF can jointly render the boundary-aligned image and semantic mask and use the semantic mask to edit the 3D volume via GAN inversion. We further show such 3D representation can be learned from widely available monocular image and semantic mask pairs. Moreover, we reveal that joint learning semantics and texture helps to generate finer geometry. Our experiments demonstrate that FENeRF outperforms state-of-the-art methods in various face editing tasks.

****

CoNeRF: Controllable Neural Radiance Fields

Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzci■ski, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18623-18632

We extend neural 3D representations to allow for intuitive and interpretable user control beyond novel view rendering (i.e. camera control). We allow the user to annotate which part of the scene one wishes to control with just a small number of mask annotations in the training images. Our key idea is to treat the attributes as latent variables that are regressed by the neural network given the scene encoding. This leads to a few-shot learning framework, where attributes are discovered automatically by the framework when annotations are not provided. We apply our method to various scenes with different types of controllable attributes (e.g. expression control on human faces, or state control in the movement of inanimate objects). Overall, we demonstrate, to the best of our knowledge, for the first time novel view and novel attribute re-rendering of scenes from a single video.

****

Noise2NoiseFlow: Realistic Camera Noise Modeling Without Clean Images

Ali Maleky, Shayan Kousha, Michael S. Brown, Marcus A. Brubaker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17632-17641

Image noise modeling is a long-standing problem with many applications in computer vision. Early attempts that propose simple models, such as signal-independent

additive white Gaussian noise or the heteroscedastic Gaussian noise model (a.k.a., camera noise level function) are not sufficient to learn the complex behavior of the camera sensor noise. Recently, more complex learning-based models have been proposed that yield better results in noise synthesis and downstream tasks, such as denoising. However, their dependence on supervised data (i.e., paired clean images) is a limiting factor given the challenges in producing ground-truth images. This paper proposes a framework for training a noise model and a denoiser simultaneously while relying only on pairs of noisy images rather than noisy/clean paired image data. We apply this framework to the training of the Noise Flow architecture. The noise synthesis and density estimation results show that our framework outperforms previous signal-processing-based noise models and is on par with its supervised counterpart. The trained denoiser is also shown to significantly improve upon both supervised and weakly supervised baseline denoising approaches. The results indicate that the joint training of a denoiser and a noise model yields significant improvements in the denoiser.

****

ZeroWaste Dataset: Towards Deformable Object Segmentation in Cluttered Scenes

Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, Kate Saenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21147-21157

Less than 35% of recyclable waste is being actually recycled in the US, which leads to increased soil and sea pollution and is one of the major concerns of environmental researchers as well as the common public. At the heart of the problem are the inefficiencies of the waste sorting process (separating paper, plastic, metal, glass, etc.) due to the extremely complex and cluttered nature of the waste stream. Recyclable waste detection poses a unique computer vision challenge as it requires detection of highly deformable and often translucent objects in cluttered scenes without the kind of context information usually present in human-centric datasets. This challenging computer vision task currently lacks suitable datasets or methods in the available literature. In this paper, we take a step towards computer-aided waste detection and present the first in-the-wild industrial-grade waste detection and segmentation dataset, ZeroWaste. We believe that ZeroWaste will catalyze research in object detection and semantic segmentation in extreme clutter as well as applications in the recycling domain. Our project page can be found at http://ai.bu.edu/zerowaste/.

****

Remember Intentions: Retrospective-Memory-Based Trajectory Prediction

Chenxin Xu, Weibo Mao, Wenjun Zhang, Siheng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6488-6497

To realize trajectory prediction, most previous methods adopt the parameter-based approach, which encodes all the seen past-future instance pairs into model parameters. However, in this way, the model parameters come from all seen instances, which means a huge amount of irrelevant seen instances might also involve in predicting the current situation, disturbing the performance. To provide a more explicit link between the current situation and the seen instances, we imitate the mechanism of retrospective memory in neuropsychology and propose MemoNet, an instance-based approach that predicts the movement intentions of agents by looking for similar scenarios in the training data. In MemoNet, we design a pair of memory banks to explicitly store representative instances in the training set, acting as prefrontal cortex in the neural system, and a trainable memory addresser to adaptively search a current situation with similar instances in the memory bank, acting like basal ganglia. During prediction, MemoNet recalls previous memory by using the memory addresser to index related instances in the memory bank. We further propose a two-step trajectory prediction system, where the first step is to leverage MemoNet to predict the destination and the second step is to fulfill the whole trajectory according to the predicted destinations. Experiments show that the proposed MemoNet improves the FDE by 20.3%/10.2%/28.3% from the previous best method on SDD/ETH-UCY/NBA datasets. Experiments also show that our MemoNet has the ability to trace back to specific instances during prediction, prom

oting more interpretability.
****
Measuring Compositional Consistency for Video Question Answering
Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, Maneesh Agrawala; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5046-5055
Recent video question answering benchmarks indicate that state-of-the-art models struggle to answer compositional questions. However, it remains unclear which types of compositional reasoning cause models to mispredict. Furthermore, it is difficult to discern whether models arrive at answers using compositional reasoning or by leveraging data biases. In this paper, we develop a question decomposition engine that programmatically deconstructs a compositional question into a directed acyclic graph of sub-questions. The graph is designed such that each parent question is a composition of its children. We present AGQA-Decomp, a benchmark containing 2.3M question graphs, with an average of 11.49 sub-questions per graph, and 4.55M total new sub-questions. Using question graphs, we evaluate three state-of-the-art models with a suite of novel compositional consistency metrics. We find that models either cannot reason correctly through most compositions or are reliant on incorrect reasoning to reach answers, frequently contradicting themselves or achieving high accuracies when failing at intermediate reasoning steps.
****
Category Contrast for Unsupervised Domain Adaptation in Visual Tasks
Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1203-1214
Instance contrast for unsupervised representation learning has achieved great success in recent years. In this work, we explore the idea of instance contrastive learning in unsupervised domain adaptation (UDA) and propose a novel Category Contrast technique (CaCo) that introduces semantic priors on top of instance discrimination for visual UDA tasks. By considering instance contrastive learning as a dictionary look-up operation, we construct a semantics-aware dictionary with samples from both source and target domains where each target sample is assigned a (pseudo) category label based on the category priors of source samples. This allows category contrastive learning (between target queries and the category-level dictionary) for category-discriminative yet domain-invariant feature representations: samples of the same category (from either source or target domain) are pulled closer while those of different categories are pushed apart simultaneously. Extensive UDA experiments in multiple visual tasks (e.g., segmentation, classification and detection) show that CaCo achieves superior performance as compared with state-of-the-art methods. The experiments also demonstrate that CaCo is complementary to existing UDA methods and generalizable to other learning setups such as unsupervised model adaptation, open-/partial-set adaptation etc.
****
SwapMix: Diagnosing and Regularizing the Over-Reliance on Visual Context in Visual Question Answering
Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, Alan Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5078-5088
While Visual Question Answering (VQA) has progressed rapidly, previous works raise concerns about robustness of current VQA models. In this work, we study the robustness of VQA models from a novel perspective: visual context. We suggest that the models over-rely on the visual context, i.e., irrelevant objects in the image, to make predictions. To diagnose the models' reliance on visual context and measure their robustness, we propose a simple yet effective perturbation technique, SwapMix. SwapMix perturbs the visual context by swapping features of irrelevant context objects with features from other objects in the dataset. Using SwapMix we are able to change answers to more than 45% of the questions for a representative VQA model. Additionally, we train the models with perfect sight and find that the context over-reliance highly depends on the quality of visual represe

ntations. In addition to diagnosing, SwapMix can also be applied as a data augme ntation strategy during training in order to regularize the context over-relianc e. By swapping the context object features, the model reliance on context can be suppressed effectively. Two representative VQA models are studied using SwapMix : a co-attention model MCAN and a large-scale pretrained model LXMERT. Our exper iments on the popular GQA dataset show the effectiveness of SwapMix for both dia gnosing model robustness, and regularizing the over-reliance on visual context.
****

UNIST: Unpaired Neural Implicit Shape Translation Network

Qimin Chen, Johannes Merz, Aditya Sanghi, Hooman Shayani, Ali Mahdavi-Amiri, Hao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 18614-18622

We introduce UNIST, the first deep neural implicit model for general-purpose, un paired shape-to-shape translation, in both 2D and 3D domains. Our model is built on autoencoding implicit fields, rather than point clouds which represents the state of the art. Furthermore, our translation network is trained to perform the task over a latent grid representation which combines the merits of both latent -space processing and position awareness, to not only enable drastic shape trans forms but also well preserve spatial features and fine local details for natural shape translations. With the same network architecture and only dictated by the input domain pairs, our model can learn both style-preserving content alteratio n and content-preserving style transfer. We demonstrate the generality and quali ty of the translation results, and compare them to well-known baselines. Code is available at https://qiminchen.github.io/unist/.
****

Local-Adaptive Face Recognition via Graph-Based Meta-Clustering and Regularized Adaptation

Wenbin Zhu, Chien-Yi Wang, Kuan-Lun Tseng, Shang-Hong Lai, Baoyuan Wang; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20301-20310

Due to the rising concern of data privacy, it's reasonable to assume the local c lient data can't be transferred to a centralized server, nor their associated id entity label is provided. To support continuous learning and fill the last-mile quality gap, we introduce a new problem setup called Local-Adaptive Face Recogni tion (LaFR). Leveraging the environment-specific local data after the deployment of the initial global model, LaFR aims at getting optimal performance by traini ng local-adapted models automatically and un-supervisely, as opposed to fixing t heir initial global model. We achieve this by a newly proposed embedding cluster model based on Graph Convolution Network (GCN), which is trained via meta-optim ization procedure. Compared with previous works, our meta-clustering model can g eneralize well in unseen local environments. With the pseudo identity labels fro m the clustering results, we further introduce novel regularization techniques t o improve the model adaptation performance. Extensive experiments on racial and internal sensor adaptation demonstrate that our proposed solution is more effect ive for adapting face recognition models in each specific environment. Meanwhile , we show that LaFR can further improve the global model by a simple federated a ggregation over the updated local models.
****

The DEVIL Is in the Details: A Diagnostic Evaluation Benchmark for Video Inpaint ing

Ryan Szeto, Jason J. Corso; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 21054-21063

Quantitative evaluation has increased dramatically among recent video inpainting work, but the video and mask content used to gauge performance has received rel atively little attention. Although attributes such as camera and background scen e motion inherently change the difficulty of the task and affect methods differe ntly, existing evaluation schemes fail to control for them, thereby providing mi nimal insight into inpainting failure modes. To address this gap, we propose the Diagnostic Evaluation of Video Inpainting on Landscapes (DEVIL) benchmark, whic h consists of two contributions: (i) a novel dataset of videos and masks labeled

according to several key inpainting failure modes, and (ii) an evaluation scheme that samples slices of the dataset characterized by a fixed content attribute, and scores performance on each slice according to reconstruction, realism, and temporal consistency quality. By revealing systematic changes in performance induced by particular characteristics of the input content, our challenging benchmark enables more insightful analysis into video inpainting methods and serves as an invaluable diagnostic tool for the field. Our code and data are available at github.com/MichiganCOG/devil.

****

Mutual Information-Driven Pan-Sharpening

Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, Feng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1798-1808

Pan-sharpening aims to integrate the complementary information of texture-rich PAN images and multi-spectral (MS) images to produce the texture-rich MS images. Despite the remarkable progress, existing state-of-the-art Pan-sharpening methods don't explicitly enforce the complementary information learning between two modalities of PAN and MS images. This leads to information redundancy not being handled well, which further limits the performance of these methods. To address the above issue, we propose a novel mutual information-driven Pan-sharpening framework in this paper. To be specific, we first project the PAN and MS image into modality-aware feature space independently, and then impose the mutual information minimization over them to explicitly encourage the complementary information learning. Such operation is capable of reducing the information redundancy and improving the model performance. Extensive experimental results over multiple satellite datasets demonstrate that the proposed algorithm outperforms other state-of-the-art methods qualitatively and quantitatively with great generalization ability to real-world scenes.

****

Shifting More Attention to Visual Backbone: Query-Modulated Refinement Networks for End-to-End Visual Grounding

Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, Xin Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15502-15512

Visual grounding focuses on establishing fine-grained alignment between vision and natural language, which has essential applications in multimodal reasoning systems. Existing methods use pre-trained query-agnostic visual backbones to extract visual feature maps independently without considering the query information. We argue that the visual features extracted from the visual backbones and the features really needed for multimodal reasoning are inconsistent. One reason is that there are differences between pre-training tasks and visual grounding. Moreover, since the backbones are query-agnostic, it is difficult to completely avoid the inconsistency issue by training the visual backbone end-to-end in the visual grounding framework. In this paper, we propose a Query-modulated Refinement Network (QRNet) to address the inconsistent issue by adjusting intermediate features in the visual backbone with a novel Query-aware Dynamic Attention (QD-ATT) mechanism and query-aware multiscale fusion. The QD-ATT can dynamically compute query-dependent visual attention at the spatial and channel level of the feature maps produced by the visual backbone. We apply the QRNet to an end-to-end visual grounding framework. Extensive experiments show that the proposed method outperforms state-of-the-art methods on five widely used datasets.

****

A Framework for Learning Ante-Hoc Explainable Models via Concepts

Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10286-10295

Self-explaining deep models are designed to learn the latent concept-based explanations implicitly during training, which eliminates the requirement of any post-hoc explanation generation technique. In this work, we propose one such model that appends an explanation generation module on top of any basic network and joi

ntly trains the whole module that shows high predictive performance and generate s meaningful explanations in terms of concepts. Our training strategy is suitabl e for unsupervised concept learning with much lesser parameter space requirement s compared to baseline methods. Our proposed model also has provision for levera ging self-supervision on concepts to extract better explanations. However, with full concept supervision, we achieve the best predictive performance compared to recently proposed concept-based explainable models. We report both qualitative and quantitative results with our method, which shows better performance than re cently proposed concept-based explainability methods. We reported exhaustive res ults with two datasets without ground truth concepts, i.e., CIFAR10, ImageNet, a nd two datasets with ground truth concepts, i.e., AwA2, CUB-200, to show the eff ectiveness of our method for both cases. To the best of our knowledge, we are th e first ante-hoc explanation generation method to show results with a large-scal e dataset such as ImageNet.

****

Generating Useful Accident-Prone Driving Scenarios via a Learned Traffic Prior

Davis Rempe, Jonah Philion, Leonidas J. Guibas, Sanja Fidler, Or Litany; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17305-17315

Evaluating and improving planning for autonomous vehicles requires scalable gene ration of long-tail traffic scenarios. To be useful, these scenarios must be rea listic and challenging, but not impossible to drive through safely. In this work , we introduce STRIVE, a method to automatically generate challenging scenarios that cause a given planner to produce undesirable behavior, like collisions. To maintain scenario plausibility, the key idea is to leverage a learned model of t raffic motion in the form of a graph-based conditional VAE. Scenario generation is formulated as an optimization in the latent space of this traffic model, pert urbing an initial real-world scene to produce trajectories that collide with a g iven planner. A subsequent optimization is used to find a "solution" to the scen ario, ensuring it is useful to improve the given planner. Further analysis clust ers generated scenarios based on collision type. We attack two planners and show that STRIVE successfully generates realistic, challenging scenarios in both cas es. We additionally "close the loop" and use these scenarios to optimize hyperpa rameters of a rule-based planner.

****

FLOAT: Factorized Learning of Object Attributes for Improved Multi-Object Multi-Part Scene Parsing

Rishubh Singh, Pranav Gupta, Pradeep Shenoy, Ravikiran Sarvadevabhatla; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2022, pp. 1445-1455

Multi-object multi-part scene parsing is a challenging task which requires detec ting multiple object classes in a scene and segmenting the semantic parts within each object. In this paper, we propose FLOAT, a factorized label space framewor k for scalable multi-object multi-part parsing. Our framework involves independe nt dense prediction of object category and part attributes which increases scala bility and reduces task complexity compared to the monolithic label space counte rpart. In addition, we propose an inference-time 'zoom' refinement technique whi ch significantly improves segmentation quality, especially for smaller objects/p arts. Compared to state of the art, FLOAT obtains an absolute improvement of 2.0 % for mean IOU (mIOU) and 4.8% for segmentation quality IOU (sqIOU) on the Pasca l-Part-58 dataset. For the larger Pascal-Part-108 dataset, the improvements are 2.1% for mIOU and 3.9% for sqIOU. We incorporate previously excluded part attrib utes and other minor parts of the Pascal-Part dataset to create the most compreh ensive and challenging version which we dub Pascal-Part-201. FLOAT obtains impro vements of 8.6% for mIOU and 7.5% for sqIOU on the new dataset, demonstrating it s parsing effectiveness across a challenging diversity of objects and parts. The code and datasets are available at floatseg.github.io.

****

Efficient Geometry-Aware 3D Generative Adversarial Networks

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini D

e Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16123-16133
Unsupervised generation of high-quality multi-view-consistent images and 3D shapes using only collections of single-view 2D photographs has been a long-standing challenge. Existing 3D GANs are either compute-intensive or make approximations that are not 3D-consistent; the former limits quality and resolution of the generated images and the latter adversely affects multi-view consistency and shape quality. In this work, we improve the computational efficiency and image quality of 3D GANs without overly relying on these approximations. We introduce an expressive hybrid explicit-implicit network architecture that, together with other design choices, synthesizes not only high-resolution multi-view-consistent images in real time but also produces high-quality 3D geometry. By decoupling feature generation and neural rendering, our framework is able to leverage state-of-the-art 2D CNN generators, such as StyleGAN2, and inherit their efficiency and expressiveness. We demonstrate state-of-the-art 3D-aware synthesis with FFHQ and AFHQ Cats, among other experiments.
****

DO-GAN: A Double Oracle Framework for Generative Adversarial Networks
Aye Phyu Phyu Aung, Xinrun Wang, Runsheng Yu, Bo An, Senthilnath Jayavelu, Xiaoli Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11275-11284
In this paper, we propose a new approach to train Generative Adversarial Networks (GANs) where we deploy a double-oracle framework using the generator and discriminator oracles. GAN is essentially a two-player zero-sum game between the generator and the discriminator. Training GANs is challenging as a pure Nash equilibrium may not exist and even finding the mixed Nash equilibrium is difficult as GANs have a large-scale strategy space. In DO-GAN, we extend the double oracle framework to GANs. We first generalize the players' strategies as the trained models of generator and discriminator from the best response oracles. We then compute the meta-strategies using a linear program. For scalability of the framework where multiple generators and discriminator best responses are stored in the memory, we propose two solutions: 1) pruning the weakly-dominated players' strategies to keep the oracles from becoming intractable; 2) applying continual learning to retain the previous knowledge of the networks. We apply our framework to established GAN architectures such as vanilla GAN, Deep Convolutional GAN, Spectral Normalization GAN and Stacked GAN. Finally, we conduct experiments on MNIST, CIFAR-10 and CelebA datasets and show that DO-GAN variants have significant improvements in both subjective qualitative evaluation and quantitative metrics, compared with their respective GAN architectures.
****

Dancing Under the Stars: Video Denoising in Starlight
Kristina Monakhova, Stephan R. Richter, Laura Waller, Vladlen Koltun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16241-16251
Imaging in low light is extremely challenging due to low photon counts. Using sensitive CMOS cameras, it is currently possible to take videos at night under moonlight (0.05-0.3 lux illumination). In this paper, we demonstrate photorealistic video under starlight (no moon present, <0.001 lux) for the first time. To enable this, we develop a GAN-tuned physics-based noise model to more accurately represent camera noise at the lowest light levels. Using this noise model, we train a video denoiser using a combination of simulated noisy video clips and real noisy still images. We capture a 5-10 fps video dataset with significant motion at approximately 0.6-0.7 millilux with no active illumination. Comparing against alternative methods, we achieve improved video quality at the lowest light levels, demonstrating photorealistic video denoising in starlight for the first time.
****

FocusCut: Diving Into a Focus View in Interactive Segmentation
Zheng Lin, Zheng-Peng Duan, Zhao Zhang, Chun-Le Guo, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

2022, pp. 2637-2646
Interactive image segmentation is an essential tool in pixel-level annotation and image editing. To obtain a high-precision binary segmentation mask, users tend to add interaction clicks around the object details, such as edges and holes, for efficient refinement. Current methods regard these repair clicks as the guidance to jointly determine the global prediction. However, the global view makes the model lose focus from later clicks, and is not in line with user intentions. In this paper, we dive into the view of clicks' eyes to endow them with the decisive role in object details again. To verify the necessity of focus view, we design a simple yet effective pipeline, named FocusCut, which integrates the functions of object segmentation and local refinement. After obtaining the global prediction, it crops click-centered patches from the original image with adaptive scopes to refine the local predictions progressively. Without user perception and parameters increase, our method has achieved state-of-the-art results. Extensive experiments and visualized results demonstrate that FocusCut makes hyper-fine segmentation possible for interactive image segmentation.
****

Medial Spectral Coordinates for 3D Shape Analysis
Morteza Rezanejad, Mohammad Khodadad, Hamidreza Mahyar, Herve Lombaert, Michael Gruninger, Dirk Walther, Kaleem Siddiqi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2686-2696
In recent years there has been a resurgence of interest in our community in the shape analysis of 3D objects represented by surface meshes, their voxelized interiors, or surface point clouds. In part, this interest has been stimulated by the increased availability of RGBD cameras, and by applications of computer vision to autonomous driving, medical imaging, and robotics. In these settings, spectral coordinates have shown promise for shape representation due to their ability to incorporate both local and global shape properties in a manner that is qualitatively invariant to isometric transformations. Yet, surprisingly, such coordinates have thus far typically considered only local surface positional or derivative information. In the present article, we propose to equip spectral coordinates with medial (object width) information, so as to enrich them. The key idea is to couple surface points that share a medial ball, via the weights of the adjacency matrix. We develop a spectral feature using this idea, and the algorithms to compute it. The incorporation of object width and medial coupling has direct benefits, as illustrated by our experiments on object classification, object part segmentation, and surface point correspondence.
****

Contextualized Spatio-Temporal Contrastive Learning With Self-Supervision
Liangzhe Yuan, Rui Qian, Yin Cui, Boqing Gong, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, Ting Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13977-13986
Modern self-supervised learning algorithms typically enforce persistency of instance representations across views. While being very effective on learning holistic image and video representations, such an objective becomes suboptimal for learning spatio-temporally fine-grained features in videos, where scenes and instances evolve through space and time. In this paper, we present Contextualized Spatio-Temporal Contrastive Learning (ConST-CL) to effectively learn spatio-temporally fine-grained video representations via self-supervision. We first design a region-based pretext task which requires the model to transform instance representations from one view to another, guided by context features. Further, we introduce a simple network design that successfully reconciles the simultaneous learning process of both holistic and local representations. We evaluate our learned representations on a variety of downstream tasks and show that ConST-CL achieves competitive results on 6 datasets, including Kinetics, UCF, HMDB, AVAKinetics, AVA and OTB. Our code and models will be available.
****

Rethinking Architecture Design for Tackling Data Heterogeneity in Federated Learning
Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli,

Li Fei-Fei, Daniel Rubin; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2022, pp. 10061-10071
Federated learning is an emerging research paradigm enabling collaborative train
ing of machine learning models among different organizations while keeping data
private at each institution. Despite recent progress, there remain fundamental c
hallenges such as the lack of convergence and the potential for catastrophic for
getting across real-world heterogeneous devices. In this paper, we demonstrate t
hat self-attention-based architectures (e.g., Transformers) are more robust to d
istribution shifts and hence improve federated learning over heterogeneous data.
Concretely, we conduct the first rigorous empirical investigation of different
neural architectures across a range of federated algorithms, real-world benchmar
ks, and heterogeneous data splits. Our experiments show that simply replacing co
nvolutional networks with Transformers can greatly reduce catastrophic forgettin
g of previous devices, accelerate convergence, and reach a better global model,
especially when dealing with heterogeneous data. We will release our code and pr
etrained models to encourage future exploration in robust architectures as an al
ternative to current research efforts on the optimization front.
****

APES: Articulated Part Extraction From Sprite Sheets
Zhan Xu, Matthew Fisher, Yang Zhou, Deepali Aneja, Rushikesh Dudhat, Li Yi, Evan
gelos Kalogerakis; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2022, pp. 11635-11644
Rigged puppets are one of the most prevalent representations to create 2D charac
ter animations. Creating these puppets requires partitioning characters into ind
ependently moving parts. In this work, we present a method to automatically iden
tify such articulated parts from a small set of character poses shown in a sprit
e sheet, which is an illustration of the character that artists often draw befor
e puppet creation. Our method is trained to infer articulated parts, e.g. head,
torso and limbs, that can be re-assembled to best reconstruct the given poses. O
ur results demonstrate significantly better performance than alternatives qualit
atively and quantitatively. Our project page https://zhan-xu.github.io/parts/ in
cludes our code and data.
****

Dressing in the Wild by Watching Dance Videos
Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K. Du, Min Zheng, Xiang Lo
ng, Xiaodan Liang, Jianchao Yang; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2022, pp. 3480-3489
While significant progress has been made in garment transfer, one of the most ap
plicable directions of human-centric image generation, existing works overlook t
he in-the-wild imagery, presenting severe garment-person misalignment as well as
noticeable degradation in fine texture details. This paper, therefore, attends
to virtual try-on in real-world scenes and brings essential improvements in auth
enticity and naturalness especially for loose garment (e.g., skirts, formal dres
ses), challenging poses (e.g., cross arms, bent legs), and cluttered backgrounds
. Specifically, we find that the pixel flow excels at handling loose garments wh
ereas the vertex flow is preferred for hard poses, and by combining their advant
ages we propose a novel generative network called wFlow that can effectively pus
h up garment transfer to in-the-wild context. Moreover, former approaches requir
e paired images for training. Instead, we cut down the laboriousness by working
on a newly constructed large-scale video dataset named Dance50k with self-superv
ised cross-frame training and an online cycle optimization. The proposed Dance50
k can boost real-world virtual dressing by covering a wide variety of garments u
nder dancing poses. Extensive experiments demonstrate the superiority of our wFl
ow in generating realistic garment transfer results for in-the-wild images witho
ut resorting to expensive paired datasets.
****

SPAct: Self-Supervised Privacy Preservation for Action Recognition
Ishan Rajendrakumar Dave, Chen Chen, Mubarak Shah; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20164-201
73

Visual private information leakage is an emerging key issue for the fast growing applications of video understanding like activity recognition. Existing approaches for mitigating privacy leakage in action recognition require privacy labels along with the action labels from the video dataset. However, annotating frames of video dataset for privacy labels is not feasible. Recent developments of self-supervised learning (SSL) have unleashed the untapped potential of the unlabeled data. For the first time, we present a novel training framework which removes privacy information from input video in a self-supervised manner without requiring privacy labels. Our training framework consists of three main components: anonymization function, self-supervised privacy removal branch, and action recognition branch. We train our framework using a minimax optimization strategy to minimize the action recognition cost function and maximize the privacy cost function through a contrastive self-supervised loss. Employing existing protocols of known-action and privacy attributes, our framework achieves a competitive action-privacy trade-off to the existing state-of-the-art supervised methods. In addition, we introduce a new protocol to evaluate the generalization of learned the anonymization function to novel-action and privacy attributes and show that our self-supervised framework outperforms existing supervised methods. Code available at: https://github.com/DAVEISHAN/SPAct

****

Uni6D: A Unified CNN Framework Without Projection Breakdown for 6D Pose Estimation

Xiaoke Jiang, Donghai Li, Hao Chen, Ye Zheng, Rui Zhao, Liwei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11174-11184

As RGB-D sensors become more affordable, using RGB-D images to obtain high-accuracy 6D pose estimation results becomes a better option. State-of-the-art approaches typically use different backbones to extract features for RGB and depth images. They use a 2D CNN for RGB images and a per-pixel point cloud network for depth data, as well as a fusion network for feature fusion. We find that the essential reason for using two independent backbones is the "projection breakdown" problem. In the depth image plane, the projected 3D structure of the physical world is preserved by the 1D depth value and its built-in 2D pixel coordinate (UV). Any spatial transformation that modifies UV, such as resize, flip, crop, or pooling operations in the CNN pipeline, breaks the binding between the pixel value and UV coordinate. As a consequence, the 3D structure is no longer preserved by a modified depth image or feature. To address this issue, we propose a simple yet effective method denoted as Uni6D that explicitly takes the extra UV data along with RGB-D images as input. Our method has a Unified CNN framework for 6D pose estimation with a single CNN backbone. In particular, the architecture of our method is based on Mask R-CNN with two extra heads, one named RT head for directly predicting 6D pose and the other named abc head for guiding the network to map the visible points to their coordinates in the 3D model as an auxiliary module. This end-to-end approach balances simplicity and accuracy, achieving comparable accuracy with state of the arts and 7.2x faster inference speed on the YCB-Video dataset.

****

De-Rendering 3D Objects in the Wild

Felix Wimbauer, Shangzhe Wu, Christian Rupprecht; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18490-18499

With increasing focus on augmented and virtual reality applications (XR) comes the demand for algorithms that can lift objects from images and videos into representations that are suitable for a wide variety of related 3D tasks. Large-scale deployment of XR devices and applications means that we cannot solely rely on supervised learning, as collecting and annotating data for the unlimited variety of objects in the real world is infeasible. We present a weakly supervised method that is able to decompose a single image of an object into shape (depth and normals), material (albedo, reflectivity and shininess) and global lighting parameters. For training, the method only relies on a rough initial shape estimate of

the training objects to bootstrap the learning process. This shape supervision can come for example from a pretrained depth network or - more generically - from a traditional structure-from-motion pipeline. In our experiments, we show that the method can successfully de-render 2D images into a decomposed 3D representation and generalizes to unseen object categories. Since in-the-wild evaluation is difficult due to the lack of ground truth data, we also introduce a photo-realistic synthetic test set that allows for quantitative evaluation. Please find our project page at: https://github.com/Brummi/derender3d
****

SPAMs: Structured Implicit Parametric Models
Pablo Palafox, Nikolaos Sarafianos, Tony Tung, Angela Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12851-12860
Parametric 3D models have formed a fundamental role in modeling deformable objects, such as human bodies, faces, and hands; however, the construction of such parametric models requires significant manual intervention and domain expertise. Recently, neural implicit 3D representations have shown great expressibility in capturing 3D shape geometry. We observe that deformable object motion is often semantically structured, and thus propose to learn Structured-implicit PArametric Models (SPAMs) as a deformable object representation that structurally decomposes non-rigid object motion into part-based disentangled representations of shape and pose, with each being represented by deep implicit functions. This enables a structured characterization of object movement, with part decomposition characterizing a lower-dimensional space in which we can establish coarse motion correspondence. In particular, we can leverage the part decompositions at test time to fit to new depth sequences of unobserved shapes, by establishing part correspondences between the input observation and our learned part spaces; this guides a robust joint optimization between the shape and pose of all parts, even under dramatic motion sequences. Experiments demonstrate that our part-aware shape and pose understanding lead to state-of-the-art performance in reconstruction and tracking of depth sequences of complex deforming object motion.
****

Global Sensing and Measurements Reuse for Image Compressed Sensing
Zi-En Fan, Feng Lian, Jia-Ni Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8954-8963
Recently, deep network-based image compressed sensing methods achieved high reconstruction quality and reduced computational overhead compared with traditional methods. However, existing methods obtain measurements only from partial features in the network and use it only once for image reconstruction. They ignore there are low, mid, and high-level features in the network and all of them are essential for high-quality reconstruction. Moreover, using measurements only once may not be enough for extracting richer information from measurements. To address these issues, we propose a novel Measurements Reuse Convolutional Compressed Sensing Network (MR-CCSNet) which employs Global Sensing Module (GSM) to collect all level features for achieving an efficient sensing and Measurements Reuse Block (MRB) to reuse measurements multiple times on multi-scale. Finally, we conduct a series of experiments on three benchmark datasets to show that our model can significantly outperform state-of-the-art methods. Code is available at https://github.com/fze0012/MR-CCSNet.
****

SeeThroughNet: Resurrection of Auxiliary Loss by Preserving Class Probability Information
Dasol Han, Jaewook Yoo, Dokwan Oh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4463-4472
Auxiliary loss is additional loss besides the main branch loss to help optimize the learning process of neural networks. In order to calculate the auxiliary loss between the feature maps of intermediate layers and the ground truth in the field of semantic segmentation, the size of each feature map must match the ground truth. In all studies using the auxiliary losses with the segmentation models, from what we have investigated, they either use a down-sampling function to redu

ce the size of the ground truth or use an up-sampling function to increase the s ize of the feature map in order to match the resolution between the feature map and the ground truth. However, in the process of selecting representative values through down-sampling and up-sampling, information loss is inevitable. In this paper, we introduce Class Probability Preserving (CPP) pooling to alleviate info rmation loss in down-sampling the ground truth in semantic segmentation tasks. W e demonstrated the superiority of the proposed method on Cityscapes, Pascal VOC, Pascal Context, and NYU-Depth-v2 datasets by using CPP pooling with auxiliary l osses based on seven popular segmentation models. In addition, we propose See-Th rough Network (SeeThroughNet) that adopts an improved multi-scale attention-coup led decoder structure to maximize the effect of CPP pooling. SeeThroughNet shows cutting-edge results in the field of semantic understanding of urban street sce nes, which ranked #1 on the Cityscapes benchmark.

****

Representing 3D Shapes With Probabilistic Directed Distance Fields

Tristan Aumentado-Armstrong, Stavros Tsogkas, Sven Dickinson, Allan D. Jepson; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19343-19354

Differentiable rendering is an essential operation in modern vision, allowing in verse graphics approaches to 3D understanding to be utilized in modern machine l earning frameworks. Yet, explicit shape representations (e.g., voxels, point clo uds, meshes), while relatively easily rendered, often suffer from limited geomet ric fidelity or topological constraints. On the other hand, implicit representat ions (e.g., occupancy, distance, or radiance fields) preserve greater fidelity, but suffer from complex or inefficient rendering processes, limiting scalability . In this work, we endeavour to address both shortcomings with a novel shape rep resentation that allows fast differentiable rendering within an implicit archite cture. Building on implicit distance representations, we define Directed Distanc e Fields (DDFs), which map an oriented point (position and direction) to surface visibility and depth. Such a field can render a depth map with a single forward pass per pixel, enable differential surface geometry extraction (e.g., surface normals and curvatures) via network derivatives, can be easily composed, and per mit extraction of classical unsigned distance fields. Using probabilistic DDFs (PDDFs), we show how to model inherent discontinuities in the underlying field. F inally, we apply our method to fitting single shapes, unpaired 3D-aware generati ve image modelling, and single-image 3D reconstruction tasks, showcasing strong performance with simple architectural components via the versatility of our repr esentation.

****

Learning ABCs: Approximate Bijective Correspondence for Isolating Factors of Var iation With Weak Supervision

Kieran A. Murphy, Varun Jampani, Srikumar Ramalingam, Ameesh Makadia; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16010-16020

Representational learning forms the backbone of most deep learning applications, and the value of a learned representation is intimately tied to its information content regarding different factors of variation. Finding good representations depends on the nature of supervision and the learning algorithm. We propose a no vel algorithm that utilizes a weak form of supervision where the data is partiti oned into sets according to certain inactive (common) factors of variation which are invariant across elements of each set. Our key insight is that by seeking c orrespondence between elements of different sets, we learn strong representation s that exclude the inactive factors of variation and isolate the active factors that vary within all sets. As a consequence of focusing on the active factors, o ur method can leverage a mix of set-supervised and wholly unsupervised data, whi ch can even belong to a different domain. We tackle the challenging problem of s ynthetic-to-real object pose transfer, without pose annotations on anything, by isolating pose information which generalizes to the category level and across th e synthetic/real domain gap. The method can also boost performance in supervised settings, by strengthening intermediate representations, as well as operate in

practically attainable scenarios with set-supervised natural images, where quant
ity is limited and nuisance factors of variation are more plentiful.
****

ABO: Dataset and Benchmarks for Real-World 3D Object Understanding
Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gun
dogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matth
ieu Guillaumin, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2022, pp. 21126-21136
We introduce Amazon Berkeley Objects (ABO), a new large-scale dataset designed t
o help bridge the gap between real and virtual 3D worlds. ABO contains product c
atalog images, metadata, and artist-created 3D models with complex geometries an
d physically-based materials that correspond to real, household objects. We deri
ve challenging benchmarks that exploit the unique properties of ABO and measure
the current limits of the state-of-the-art on three open problems for real-world
 3D object understanding: single-view 3D reconstruction, material estimation, an
d cross-domain multi-view object retrieval.
****

DETReg: Unsupervised Pretraining With Region Priors for Object Detection
Amir Bar, Xin Wang, Vadim Kantorov, Colorado J. Reed, Roei Herzig, Gal Chechik,
Anna Rohrbach, Trevor Darrell, Amir Globerson; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14605-14615
Recent self-supervised pretraining methods for object detection largely focus on
 pretraining the backbone of the object detector, neglecting key parts of detect
ion architecture. Instead, we introduce DETReg, a new self-supervised method tha
t pretrains the entire object detection network, including the object localizati
on and embedding components. During pretraining, DETReg predicts object localiza
tions to match the localizations from an unsupervised region proposal generator
and simultaneously aligns the corresponding feature embeddings with embeddings f
rom a self-supervised image encoder. We implement DETReg using the DETR family o
f detectors and show that it improves over competitive baselines when finetuned
on COCO, PASCAL VOC, and Airbus Ship benchmarks. In low-data regimes, including
semi-supervised and few-shot learning settings, DETReg establishes many state-of
-the-art results, e.g., on COCO we see a +6.0 AP improvement for 10-shot detecti
on and +3.5 AP improvement when training with only 1% of the labels.
****

Learning To Restore 3D Face From In-the-Wild Degraded Images
Zhenyu Zhang, Yanhao Ge, Ying Tai, Xiaoming Huang, Chengjie Wang, Hao Tang, Dong
jin Huang, Zhifeng Xie; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2022, pp. 4237-4247
In-the-wild 3D face modelling is a challenging problem as the predicted facial g
eometry and texture suffer from a lack of reliable clues or priors, when the inp
ut images are degraded. To address such a problem, in this paper we propose a no
vel Learning to Restore (L2R) 3D face framework for unsupervised high-quality fa
ce reconstruction from low-resolution images. Rather than directly refining 2D i
mage appearance, L2R learns to recover fine-grained 3D details on the proxy agai
nst degradation via extracting generative facial priors. Concretely, L2R propose
s a novel albedo restoration network to model high-quality 3D facial texture, in
 which the diverse guidance from the pre-trained Generative Adversarial Networks
 (GANs) is leveraged to complement the lack of input facial clues. With the fine
r details of the restored 3D texture, L2R then learns displacement maps from scr
atch to enhance the significant facial structure and geometry. Both of the proce
dures are mutually optimized with a novel 3D-aware adversarial loss, which furth
er improves the modelling performance and suppresses the potential uncertainty.
Extensive experiments on benchmarks show that L2R outperforms state-of-the-art m
ethods under the condition of low-quality inputs, and obtains superior performan
ces than 2D pre-processed modelling approaches with limited 3D proxy.
****

Practical Evaluation of Adversarial Robustness via Adaptive Auto Attack
Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, Jingkuan Song; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C

VPR), 2022, pp. 15105-15114
Defense models against adversarial attacks have grown significantly, but the lac k of practical evaluation methods has hindered progress. Evaluation can be defin ed as looking for defense models' lower bound of robustness given a budget numbe r of iterations and a test dataset. A practical evaluation method should be conv enient (i.e., parameter-free), efficient (i.e., fewer iterations) and reliable (i.e., approaching the lower bound of robustness). Towards this target, we propos e a parameter-free Adaptive Auto Attack (A3) evaluation method which addresses t he efficiency and reliability in a test-time-training fashion. Specifically, by observing that adversarial examples to a specific defense model follow some regu larities in their starting points, we design an Adaptive Direction Initializatio n strategy to speed up the evaluation. Furthermore, to approach the lower bound of robustness under the budget number of iterations, we propose an online statis tics-based discarding strategy that automatically identifies and abandons hard-t o-attack images. Extensive experiments on nearly 50 widely-used defense models d emonstrate the effectiveness of our A3.By consuming much fewer iterations than e xisting methods, i.e., 1/10 on average (10x speed up), we achieve lower robust a ccuracy in all cases. Notably, we won first place out of 1681 teams in CVPR 2021 White-box Adversarial Attacks on Defense Models competitions with this method. Code is available at: https://github.com/liuye6666/adaptive_auto_attack
****

Convolutions for Spatial Interaction Modeling
Zhaoen Su, Chao Wang, David Bradley, Carlos Vallespi-Gonzalez, Carl Wellington, Nemanja Djuric; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2022, pp. 6583-6592
In many different fields interactions between objects play a critical role in de termining their behavior. Graph neural networks (GNNs) have emerged as a powerfu l tool for modeling interactions, although often at the cost of adding considera ble complexity and latency. In this paper, we consider the problem of spatial in teraction modeling in the context of predicting the motion of actors around auto nomous vehicles, and investigate alternatives to GNNs. We revisit 2D convolution s and show that they can demonstrate comparable performance to graph networks in modeling spatial interactions with lower latency, thus providing an effective a nd efficient alternative in time-critical systems. Moreover, we propose a novel interaction loss to further improve the interaction modeling of the considered m ethods.
****

MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection
Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S. Ryoo, François Brémond; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20041-20051
Action detection is an essential and challenging task, especially for densely la belled datasets of untrimmed videos. The temporal relation is complex in those d atasets, including challenges like composite action, and co-occurring action. Fo r detecting actions in those complex videos, efficiently capturing both short-te rm and long-term temporal information in the video is critical. To this end, we propose a novel ConvTransformer network for action detection. This network compr ises three main components: (1) Temporal Encoder module extensively explores glo bal and local temporal relations at multiple temporal resolutions. (2) Temporal Scale Mixer module effectively fuses the multi-scale features to have a unified feature representation. (3) Classification module is used to learn the instance center-relative position and predict the frame-level classification scores. The extensive experiments on multiple datasets, including Charades, TSU and MultiTHU MOS, confirm the effectiveness of our proposed method. Our network outperforms t he state-of-the-art methods on all the three datasets.
****

Salvage of Supervision in Weakly Supervised Object Detection
Lin Sui, Chen-Lin Zhang, Jianxin Wu; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2022, pp. 14227-14236
Weakly supervised object detection (WSOD) has recently attracted much attention.

However, the lack of bounding-box supervision makes its accuracy much lower than fully supervised object detection (FSOD), and currently modern FSOD techniques cannot be applied to WSOD. To bridge the performance and technical gaps between WSOD and FSOD, this paper proposes a new framework, Salvage of Supervision (SoS), with the key idea being to harness every potentially useful supervisory signal in WSOD: the weak image-level labels, the pseudo-labels, and the power of semi-supervised object detection. This paper shows that each type of supervisory signal brings in notable improvements, outperforms existing WSOD methods (which mainly use only the weak labels) by large margins. The proposed SoS-WSOD method also have the ability to freely use modern FSOD techniques. SoS-WSOD achieves 64.4 mAP50 on VOC2007, 61.9 mAP50 on VOC2012 and 16.6 mAP50:95 on MS-COCO, and also has fast inference speed. Ablations and visualization further verify the effectiveness of SoS.

****

Cross-View Transformers for Real-Time Map-View Semantic Segmentation

Brady Zhou, Philipp Krähenbühl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13760-13769

We present cross-view transformers, an efficient attention-based model for map-view semantic segmentation from multiple cameras. Our architecture implicitly learns a mapping from individual camera views into a canonical map-view representation using a camera-aware cross-view attention mechanism. Each camera uses positional embeddings that depend on its intrinsic and extrinsic calibration. These embeddings allow a transformer to learn the mapping across different views without ever explicitly modeling it geometrically. The architecture consists of a convolutional image encoder for each view and cross-view transformer layers to infer a map-view semantic segmentation. Our model is simple, easily parallelizable, and runs in real-time. The presented architecture performs at state-of-the-art on the nuScenes dataset, with 4x faster inference speeds. Code is available at https://github.com/bradyz/cross_view_transformers.

****

Distinguishing Unseen From Seen for Generalized Zero-Shot Learning

Hongzu Su, Jingjing Li, Zhi Chen, Lei Zhu, Ke Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7885-7894

Generalized zero-shot learning (GZSL) aims to recognize samples whose categories may not have been seen at training. Recognizing unseen classes as seen ones or vice versa often leads to poor performance in GZSL. Therefore, distinguishing seen and unseen domains is naturally an effective yet challenging solution for GZSL. In this paper, we present a novel method which leverages both visual and semantic modalities to distinguish seen and unseen categories. Specifically, our method deploys two variational autoencoders to generate latent representations for visual and semantic modalities in a shared latent space, in which we align latent representations of both modalities by Wasserstein distance and reconstruct two modalities with the representations of each other. In order to learn a clearer boundary between seen and unseen classes, we propose a two-stage training strategy which takes advantage of seen and unseen semantic descriptions and searches a threshold to separate seen and unseen visual samples. At last, a seen expert and an unseen expert are used for final classification. Extensive experiments on five widely used benchmarks verify that the proposed method can significantly improve the results of GZSL. For instance, our method correctly recognizes more than 99% samples when separating domains and improves the final classification accuracy from 72.6% to 82.9% on AWA1.

****

Online Continual Learning on a Contaminated Data Stream With Blurry Task Boundaries

Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, Jonghyun Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9275-9284

Learning under a continuously changing data distribution with incorrect labels is a desirable real-world problem yet challenging. Large body of continual learning (CL) methods, however, assumes data streams with clean labels, and online lea

rning scenarios under noisy data streams are yet underexplored. We consider a mo re practical CL setup of an online learning from blurry data stream with corrupt ed noise, where existing CL methods struggle. To address the task, we first argu e the importance of both diversity and purity of examples in the episodic memory of continual learning models. To balance diversity and purity in the episodic m emory, we propose a novel strategy to manage and use the memory by a unified app roach of label noise aware diverse sampling and robust learning with semi-superv ised learning. Our empirical validations on four real-world or synthetic benchma rk datasets (CIFAR10 and 100, mini-WebVision, and Food-101N) show that our metho d significantly outperforms prior arts in this realistic and challenging continu al learning scenario.

****

Controllable Dynamic Multi-Task Architectures

Dripta S. Raychaudhuri, Yumin Suh, Samuel Schulter, Xiang Yu, Masoud Faraki, Ami t K. Roy-Chowdhury, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10955-10964

Multi-task learning commonly encounters competition for resources among tasks, s pecifically when model capacity is limited. This challenge motivates models whic h allow control over the relative importance of tasks and total compute cost dur ing inference time. In this work, we propose such a controllable multi-task netw ork that dynamically adjusts its architecture and weights to match the desired t ask preference as well as the resource constraints. In contrast to the existing dynamic multi-task approaches that adjust only the weights within a fixed archit ecture, our approach affords the flexibility to dynamically control the total co mputational cost and match the user-preferred task importance better. We propose a disentangled training of two hypernetworks, by exploiting task affinity and a novel branching regularized loss, to take input preferences and accordingly pre dict tree-structured models with adapted weights. Experiments on three multi-tas k benchmarks, namely PASCAL-Context, NYU-v2, and CIFAR-100, show the efficacy of our approach. Project page is available at https://www.nec-labs.com/ mas/DYMU.

****

Learning To Imagine: Diversify Memory for Incremental Learning Using Unlabeled D ata

Yu-Ming Tang, Yi-Xing Peng, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9549-9558

Deep neural network (DNN) suffers from catastrophic forgetting when learning inc rementally, which greatly limits its applications. Although maintaining a handfu l of samples (called "exemplars") of each task could alleviate forgetting to som e extent, existing methods are still limited by the small number of exemplars si nce these exemplars are too few to carry enough task-specific knowledge, and the refore the forgetting remains. To overcome this problem, we propose to "imagine" diverse counterparts of given exemplars referring to the abundant semantic-irre levant information from unlabeled data. Specifically, we develop a learnable fea ture generator to diversify exemplars by adaptively generating diverse counterpa rts of exemplars based on semantic information from exemplars and semantically-i rrelevant information from unlabeled data. We introduce semantic contrastive lea rning to enforce the generated samples to be semantic consistent with exemplars and perform semanticdecoupling contrastive learning to encourage diversity of ge nerated samples. The diverse generated samples could effectively prevent DNN fro m forgetting when learning new tasks. Our method does not bring any extra infere nce cost and outperforms state-of-the-art methods on two benchmarks CIFAR-100 an d ImageNet-Subset by a clear margin.

****

SmartAdapt: Multi-Branch Object Detection Framework for Videos on Mobiles

Ran Xu, Fangzhou Mu, Jayoung Lee, Preeti Mukherjee, Somali Chaterji, Saurabh Bag chi, Yin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 2528-2538

Several recent works seek to create lightweight deep networks for video object d etection on mobiles. We observe that many existing detectors, previously deemed computationally costly for mobiles, intrinsically support adaptive inference, an

d offer a multi-branch object detection framework (MBODF). Here, an MBODF is ref
erred to as a solution that has many execution branches and one can dynamically
choose from among them at inference time to satisfy varying latency requirements
(e.g. by varying resolution of an input frame). In this paper, we ask, and answ
er, the wide-ranging question across all MBODFs: How to expose the right set of
execution branches and then how to schedule the optimal one at inference time? I
n addition, we uncover the importance of making a content-aware decision on whic
h branch to run, as the optimal one is conditioned on the video content. Finally
, we explore a content-aware scheduler, an Oracle one, and then a practical one,
leveraging various lightweight feature extractors. Our evaluation shows that la
yered on Faster R-CNN-based MBODF, compared to 7 baselines, our SMARTADAPT achie
ves a higher Pareto optimal curve in the accuracy-vs-latency space for the ILSVR
C VID dataset.
****

VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks
Yi-Lin Sung, Jaemin Cho, Mohit Bansal; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5227-5237
Recently, fine-tuning language models pre-trained on large text corpora have pro
vided huge improvements on vision-and-language (V&L) tasks as well as on pure la
nguage tasks. However, fine-tuning the entire parameter set of pre-trained model
s becomes impractical since the model size is growing rapidly. Hence, in this pa
per, we introduce adapter-based parameter-efficient transfer learning techniques
to V&L models such as VL-BART and VL-T5. We evaluate our methods in a unified m
ulti-task setup on both image-text and video-text benchmarks. For the image-text
tasks, we use four diverse V&L datasets: VQAv2, GQA, NLVR2, and MSCOCO image ca
ptioning. For video-text tasks, we use TVQA, How2QA, TVC, and YC2C. With careful
training and thorough experiments, we benchmark three popular adapter-based met
hods (Adapter, Hyperformer, Compacter) against the standard full fine-tuning and
the recently proposed prompt-tuning approach. We also enhance the efficiency an
d performance of adapters by sharing their weights to attain knowledge across ta
sks. Our results demonstrate that training the adapter with the weight-sharing t
echnique (4.18% of total parameters for image-text tasks and 3.39% for video-tex
t tasks) can match the performance of fine-tuning the entire model. Lastly, we p
resent a comprehensive analysis including the combination of adapter and task-sp
ecific prompts and the impact of V&L pre-training on adapters.
****

Deep Hybrid Models for Out-of-Distribution Detection
Senqi Cao, Zhongfei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2022, pp. 4733-4743
We propose a principled and practical method for out-of-distribution (OoD) detec
tion with deep hybrid models (DHMs), which model the joint density p(x,y) of fea
tures and labels with a single forward pass. By factorizing the joint density p(
x,y) into three sources of uncertainty, we show that our approach has the abilit
y to identify samples semantically different from the training data. To ensure c
omputational scalability, we add a weight normalization step during training, wh
ich enables us to plug in state-of-the-art (SoTA) deep neural network (DNN) arch
itectures for approximately modeling and inferring expressive probability distri
butions. Our method provides an efficient, general, and flexible framework for p
redictive uncertainty estimation with promising results and theoretical support.
To our knowledge, this is the first work to reach 100% in OoD detection tasks o
n both vision and language datasets, especially on notably difficult dataset pai
rs such as CIFAR-10 vs. SVHN and CIFAR-100 vs. CIFAR-10. This work is a step tow
ards enabling DNNs in real-world deployment for safety-critical applications.
****

Accelerating Video Object Segmentation With Compressed Video
Kai Xu, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2022, pp. 1342-1351
We propose an efficient plug-and-play acceleration framework for semi-supervised
video object segmentation by exploiting the temporal redundancies in videos pre
sented by the compressed bitstream. Specifically, we propose a motion vector-bas

ed warping method for propagating segmentation masks from keyframes to other fra mes in a bi-directional and multi-hop manner. Additionally, we introduce a resid ual-based correction module that can fix wrongly propagated segmentation masks f rom noisy or erroneous motion vectors. Our approach is flexible and can be added on top of several existing video object segmentation algorithms. We achieved hi ghly competitive results on DAVIS17 and YouTube-VOS on various base models with substantial speed-ups of up to 3.5X with minor drops in accuracy.
****

Exploring Domain-Invariant Parameters for Source Free Domain Adaptation
Fan Wang, Zhongyi Han, Yongshun Gong, Yilong Yin; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7151-7160
Source-free domain adaptation (SFDA) newly emerges to transfer the relevant know ledge of a well-trained source model to an unlabeled target domain, which is cri tical in various privacy-preserving scenarios. Most existing methods focus on le arning the domain-invariant representations depending solely on the target data, leading to the obtained representations are target-specific. In this way, they cannot fully address the distribution shift problem across domains. In contrast, we provide a fascinating insight: rather than attempting to learn domain-invari ant representations, it is better to explore the domain-invariant parameters of the source model. The motivation behind this insight is clear: the domain-invari ant representations are dominated by only partial parameters of an available dee p source model. We devise the Domain-Invariant Parameter Exploring (DIPE) approa ch to capture such domain-invariant parameters in the source model to generate d omain-invariant representations. A distinguishing method is developed correspond ingly for two types of parameters, i.e., domain-invariant and domain-specific pa rameters, as well as an effective update strategy based on the clustering correc tion technique and a target hypothesis is proposed. Extensive experiments verify that DIPE successfully exceeds the current state-of-the-art models on many doma in adaptation datasets.
****

FastDOG: Fast Discrete Optimization on GPU
Ahmed Abbas, Paul Swoboda; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 439-449
We present a massively parallel Lagrange decomposition method for solving 0--1 i nteger linear programs occurring in structured prediction. We propose a new iter ative update scheme for solving the Lagrangean dual and a perturbation technique for decoding primal solutions. For representing subproblems we follow Lange et al. (2021) and use binary decision diagrams (BDDs). Our primal and dual algorith ms require little synchronization between subproblems and optimization over BDDs needs only elementary operations without complicated control flow. This allows us to exploit the parallelism offered by GPUs for all components of our method. We present experimental results on combinatorial problems from MAP inference for Markov Random Fields, quadratic assignment and cell tracking for developmental biology. Our highly parallel GPU implementation improves upon the running times of the algorithms from Lange et al. (2021) by up to an order of magnitude. In pa rticular, we come close to or outperform some state-of-the-art specialized heuri stics while being problem agnostic. Our implementation is available at https://g ithub.com/LPMP/BDD.
****

Fire Together Wire Together: A Dynamic Pruning Approach With Self-Supervised Mas k Prediction
Sara Elkerdawy, Mostafa Elhoushi, Hong Zhang, Nilanjan Ray; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12454-12463
Dynamic model pruning is a recent direction that allows for the inference of a d ifferent sub-network for each input sample during deployment. However, current d ynamic methods rely on learning a continuous channel gating through regularizati on by inducing sparsity loss. This formulation introduces complexity in balancin g different losses (e.g task loss, regularization loss). In addition, regulariza tion based methods lack transparent tradeoff hyperparameter selection to realize

computational budget. Our contribution is two-fold: 1) decoupled task and pruning training. 2) Simple hyperparameter selection that enables FLOPs reduction estimation before training. Inspired by the Hebbian theory in Neuroscience: "neurons that fire together wire together", we propose to predict a mask to process k filters in a layer based on the activation of its previous layer. We pose the problem as a self-supervised binary classification problem. Each mask predictor module is trained to predict if the log-likelihood for each filter in the current layer belongs to the top-k activated filters. The value k is dynamically estimated for each input based on a novel criterion using the mass of heatmaps. We show experiments on several neural architectures, such as VGG, ResNet and MobileNet on CIFAR and ImageNet datasets. On CIFAR, we reach similar accuracy to SOTA methods with 15% and 24% higher FLOPs reduction. Similarly in ImageNet, we achieve lower drop in accuracy with up to 13% improvement in FLOPs reduction.
****

Multi-Source Uncertainty Mining for Deep Unsupervised Saliency Detection
Yifan Wang, Wenbo Zhang, Lijun Wang, Ting Liu, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11727-11736
Deep learning-based image salient object detection (SOD) heavily relies on large-scale training data with pixel-wise labeling. High-quality labels involve intensive labor and are expensive to acquire. In this paper, we propose a novel multi-source uncertainty mining method to facilitate unsupervised deep learning from multiple noisy labels generated by traditional handcrafted SOD methods. We design an Uncertainty Mining Network (UMNet) which consists of multiple Merge-and-Split (MS) modules to recursively analyze the commonality and difference among multiple noisy labels and infer pixel-wise uncertainty map for each label. Meanwhile, we model the noisy labels using Gibbs distribution and propose a weighted uncertainty loss to jointly train the UMNet with the SOD network. As a consequence, our UMNet can adaptively select reliable labels for SOD network learning. Extensive experiments on benchmark datasets demonstrate that our method not only outperforms existing unsupervised methods, but also is on par with fully-supervised state-of-the-art models.
****

Self-Supervised Equivariant Learning for Oriented Keypoint Detection
Jongmin Lee, Byungjin Kim, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4847-4857
Detecting robust keypoints from an image is an integral part of many computer vision problems, and the characteristic orientation and scale of keypoints play an important role for keypoint description and matching. Existing learning-based methods for keypoint detection rely on standard translation-equivariant CNNs but often fail to detect reliable keypoints against geometric variations. To learn to detect robust oriented keypoints, we introduce a self-supervised learning framework using rotation-equivariant CNNs. We propose a dense orientation alignment loss by an image pair generated by synthetic transformations for training a histogram-based orientation map. Our method outperforms the previous methods on an image matching benchmark and a camera pose estimation benchmark.
****

Wavelet Knowledge Distillation: Towards Efficient Image-to-Image Translation
Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, Kaisheng Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12464-12474
Remarkable achievements have been attained with Generative Adversarial Networks (GANs) in image-to-image translation. However, due to a tremendous amount of parameters, state-of-the-art GANs usually suffer from low efficiency and bulky memory usage. To tackle this challenge, firstly, this paper investigates GANs performance from a frequency perspective. The results show that GANs, especially small GANs lack the ability to generate high-quality high frequency information. To address this problem, we propose a novel knowledge distillation method referred to as wavelet knowledge distillation. Instead of directly distilling the generated images of teachers, wavelet knowledge distillation first decomposes the images

into different frequency bands with discrete wavelet transformation and then only distills the high frequency bands. As a result, the student GAN can pay more attention to its learning on high frequency bands. Experiments demonstrate that our method leads to 7.08X compression and 6.80X acceleration on CycleGAN with almost no performance drop. Additionally, we have studied the relation between discriminators and generators which shows that the compression of discriminators can promote the performance of compressed generators.
****

Focal and Global Knowledge Distillation for Detectors
Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, Chun Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4643-4652
Knowledge distillation has been applied to image classification successfully. However, object detection is much more sophisticated and most knowledge distillation methods have failed on it. In this paper, we point out that in object detection, the features of the teacher and student vary greatly in different areas, especially in the foreground and background. If we distill them equally, the uneven differences between feature maps will negatively affect the distillation. Thus, we propose Focal and Global Distillation (FGD). Focal distillation separates the foreground and background, forcing the student to focus on the teacher's critical pixels and channels. Global distillation rebuilds the relation between different pixels and transfers it from teachers to students, compensating for missing global information in focal distillation. As our method only needs to calculate the loss on the feature map, FGD can be applied to various detectors. We experiment on various detectors with different backbones and the results show that the student detector achieves excellent mAP improvement. For example, ResNet-50 based RetinaNet, Faster RCNN, RepPoints and Mask RCNN with our distillation method achieve 40.7%, 42.0%, 42.0% and 42.1% mAP on COCO2017, which are 3.3, 3.6, 3.4 and 2.9 higher than the baseline, respectively. Our codes are available at https://github.com/yzd-v/FGD.
****

Learning To Prompt for Continual Learning
Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, Tomas Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 139-149
The mainstream paradigm behind continual learning has been to adapt the model parameters to non-stationary data distributions, where catastrophic forgetting is the central challenge. Typical methods rely on a rehearsal buffer or known task identity at test time to retrieve learned knowledge and address forgetting, while this work presents a new paradigm for continual learning that aims to train a more succinct memory system without accessing task identity at test time. Our method learns to dynamically prompt (L2P) a pre-trained model to learn tasks sequentially under different task transitions. In our proposed framework, prompts are small learnable parameters, which are maintained in a memory space. The objective is to optimize prompts to instruct the model prediction and explicitly manage task-invariant and task-specific knowledge while maintaining model plasticity. We conduct comprehensive experiments under popular image classification benchmarks with different challenging continual learning settings, where L2P consistently outperforms prior state-of-the-art methods. Surprisingly, L2P achieves competitive results against rehearsal-based methods even without a rehearsal buffer and is directly applicable to challenging task-agnostic continual learning. Source code is available at https://github.com/google-research/l2p.
****

Human Mesh Recovery From Multiple Shots
Georgios Pavlakos, Jitendra Malik, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1485-1495
Videos from edited media like movies are a useful, yet under-explored source of information, with rich variety of appearance and interactions between humans depicted over a large temporal context. However, the richness of data comes at the

expense of fundamental challenges such as abrupt shot changes and close up shots of actors with heavy truncation, which limits the applicability of existing 3D human understanding methods. In this paper, we address these limitations with the insight that while shot changes of the same scene incur a discontinuity between frames, the 3D structure of the scene still changes smoothly. This allows us to handle frames before and after the shot change as multi-view signal that provide strong cues to recover the 3D state of the actors. We propose a multi-shot optimization framework that realizes this insight, leading to improved 3D reconstruction and mining of sequences with pseudo-ground truth 3D human mesh. We treat this data as valuable supervision for models that enable human mesh recovery from movies; both from single image and from video, where we propose a transformer-based temporal encoder that can naturally handle missing observations due to shot changes in the input frames. We demonstrate the importance of our insight and proposed models through extensive experiments. The tools we develop open the door to processing and analyzing in 3D content from a large library of edited media, which could be helpful for many downstream applications. Code, models and data are available at: https://geopavlakos.github.io/multishot/

****

Improving Adversarial Transferability via Neuron Attribution-Based Attacks
Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, Michael R. Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14993-15002

Deep neural networks (DNNs) are known to be vulnerable to adversarial examples. It is thus imperative to devise effective attack algorithms to identify the deficiencies of DNNs beforehand in security-sensitive applications. To efficiently tackle the black-box setting where the target model's particulars are unknown, feature-level transfer-based attacks propose to contaminate the intermediate feature outputs of local models, and then directly employ the crafted adversarial samples to attack the target model. Due to the transferability of features, feature-level attacks have shown promise in synthesizing more transferable adversarial samples. However, existing feature-level attacks generally employ inaccurate neuron importance estimations, which deteriorates their transferability. To overcome such pitfalls, in this paper, we propose the Neuron Attribution-based Attack (NAA), which conducts feature-level attacks with more accurate neuron importance estimations. Specifically, we first completely attribute a model's output to each neuron in a middle layer. We then derive an approximation scheme of neuron attribution to tremendously reduce the computation overhead. Finally, we weight neurons based on their attribution results and launch feature-level attacks. Extensive experiments confirm the superiority of our approach to the state-of-the-art benchmarks.

****

Better Trigger Inversion Optimization in Backdoor Scanning
Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, Pan Li, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13368-13378

Backdoor attacks aim to cause misclassification of a subject model by stamping a trigger to inputs. Backdoors could be injected through malicious training and naturally exist. Deriving backdoor trigger for a subject model is critical to both attack and defense. A popular trigger inversion method is by optimization. Existing methods are based on finding a smallest trigger that can uniformly flip a set of input samples by minimizing a mask. The mask defines the set of pixels that ought to be perturbed. We develop a new optimization method that directly minimizes individual pixel changes, without using a mask. Our experiments show that compared to existing methods, the new one can generate triggers that require a smaller number of input pixels to be perturbed, have a higher attack success rate, and are more robust. They are hence more desirable when used in real-world attacks and more effective when used in defense. Our method is also more cost-effective.

****

GANSeg: Learning To Segment by Unsupervised Hierarchical Image Generation

Xingzhe He, Bastian Wandt, Helge Rhodin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1225-1235
Segmenting an image into its parts is a frequent preprocess for high-level visio n tasks such as image editing. However, annotating masks for supervised training is expensive. Weakly-supervised and unsupervised methods exist, but they depend on the comparison of pairs of images, such as from multi-views, frames of video s, and image augmentation, which limits their applicability. To address this, we propose a GAN-based approach that generates images conditioned on latent masks, thereby alleviating full or weak annotations required in previous approaches. W e show that such mask-conditioned image generation can be learned faithfully whe n conditioning the masks in a hierarchical manner on latent keypoints that defin e the position of parts explicitly. Without requiring supervision of masks or po ints, this strategy increases robustness to viewpoint and object positions chang es. It also lets us generate image-mask pairs for training a segmentation networ k, which outperforms the state-of-the-art unsupervised segmentation methods on e stablished benchmarks.
****

Dense Learning Based Semi-Supervised Object Detection
Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, Xian-Sheng Hua; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 4815-4824
The ultimate goal of semi-supervised object detection (SSOD) is to facilitate th e utilization and deployment of detectors in actual applications with the help o f a large amount of unlabeled data. Although a few works have proposed various s elf-training-based methods or consistency-regularization-based methods, they all target anchor-based detectors, while ignoring the dependency on anchor-free det ectors of the actual industrial deployment. To this end, in this paper, we inten d to bridge the gap on anchor-free SSOD algorithm by proposing a DenSe Learning (DSL) based algorithm for SSOD. It is mainly achieved by introducing several nov el techniques, including (1) Adaptive Ignoring strategy with MetaNet for assigni ng multi-level and accurate dense pixel-wise pseudo-labels, (2) Aggregated Teach er for producing stable and precise pseudo-labels, and (3) uncertainty consisten cy regularization among scales and shuffled patches for improving the generaliza tion of the detector. In order to verify the effectiveness of our proposed metho d, extensive experiments have been conducted over the popular datasets MS-COCO [??] and PASCAL-VOC [??], achieving state-of-the-art performances. Codes will be available at \textcolor[rgb] 1,0,0 xxxxxxxxx .
****

Fixing Malfunctional Objects With Learned Physical Simulation and Functional Pre diction
Yining Hong, Kaichun Mo, Li Yi, Leonidas J. Guibas, Antonio Torralba, Joshua B. Tenenbaum, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1413-1423
This paper studies the problem of fixing malfunctional 3D objects. While previou s works focus on building passive perception models to learn the functionality f rom static 3D objects, we argue that functionality is reckoned with respect to t he physical interactions between the object and the user. Given a malfunctional object, humans can perform mental simulations to reason about its functionality and figure out how to fix it. Inspired by this, we propose FixIt, a dataset that contains around 5k poorly-designed 3D physical objects paired with choices to f ix them. To mimic humans' mental simulation process, we present FixNet, a novel framework that seamlessly incorporates perception and physical dynamics. Specifi cally, FixNet consists of a perception module to extract the structured represen tation from the 3D point cloud, a physical dynamics prediction module to simulat e the results of interactions on 3D objects, and a functionality prediction modu le to evaluate the functionality and choose the correct fix. Experimental result s show that our framework outperforms baseline models by a large margin, and can generalize well to objects with similar interaction types. We will release our code and dataset.
****

Convolution of Convolution: Let Kernels Spatially Collaborate
Rongzhen Zhao, Jian Li, Zhenzhi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 651-660
In the biological visual pathway, especially the retina, neurons are tiled along spatial dimensions with the electrical coupling as their local association, while in a convolution layer, kernels are placed along the channel dimension singly. We propose Convolution of Convolution, associating kernels in a layer and letting them collaborate spatially. With this method, a layer can provide feature maps with extra transformations and learn its kernels together instead of isolatedly. It is only used during training, bringing in negligible extra costs; and can be re-parameterized to common convolution before testing, boosting performance gratuitously in tasks like classification, detection and segmentation. Our method works even better when large receptive fields are demanded. The code is available on site: https://github.com/Genera1Z/ConvolutionOfConvolution.
****

Make It Move: Controllable Image-to-Video Generation With Text Descriptions
Yaosi Hu, Chong Luo, Zhenzhong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18219-18228
Generating controllable videos conforming to user intentions is an appealing yet challenging topic in computer vision. To enable maneuverable control in line with user intentions, a novel video generation task, named Text-Image-to-Video generation (TI2V), is proposed. With both controllable appearance and motion, TI2V aims at generating videos from a static image and a text description. The key challenges of TI2V task lie both in aligning appearance and motion from different modalities, and in handling uncertainty in text descriptions. To address these challenges, we propose a Motion Anchor-based video GEnerator (MAGE) with an innovative motion anchor (MA) structure to store appearance-motion aligned representation. To model the uncertainty and increase the diversity, it further allows the injection of explicit condition and implicit randomness. Through three-dimensional axial transformers, MA is interacted with given image to generate next frames recursively with satisfying controllability and diversity. Accompanying the new task, we build two new video-text paired datasets based on MNIST and CATER for evaluation. Experiments conducted on these datasets verify the effectiveness of MAGE and show appealing potentials of TI2V task. Code and datasets are released at https://github.com/Youncy-Hu/MAGE.
****

C2AM Loss: Chasing a Better Decision Boundary for Long-Tail Object Detection
Tong Wang, Yousong Zhu, Yingying Chen, Chaoyang Zhao, Bin Yu, Jinqiao Wang, Ming Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6980-6989
Long-tail object detection suffers from poor performance on tail categories. We reveal that the real culprit lies in the extremely imbalanced distribution of the classifier's weight norm. For conventional softmax cross-entropy loss, such imbalanced weight norm distribution yields ill conditioned decision boundary for categories which have small weight norms. To get rid of this situation, we choose to maximize the cosine similarity between the learned feature and the weight vector of target category rather than the inner-product of them. The decision boundary between any two categories is the angular bisector of their weight vectors. Whereas, the absolutely equal decision boundary is suboptimal because it reduces the model's sensitivity to various categories. Intuitively, categories with rich data diversity should occupy a larger area in the classification space while categories with limited data diversity should occupy a slightly small space. Hence, we devise a Category-Aware Angular Margin Loss (C2AM Loss) to introduce an adaptive angular margin between any two categories. Specifically, the margin between two categories is proportional to the ratio of their classifiers' weight norms. As a result, the decision boundary is slightly pushed towards the category which has a smaller weight norm. We conduct comprehensive experiments on LVIS dataset. C2AM Loss brings 4.9 5.2 AP improvements on different detectors and backbones compared with baseline.
****

Neural Points: Point Cloud Representation With Neural Fields for Arbitrary Upsampling
Wanquan Feng, Jin Li, Hongrui Cai, Xiaonan Luo, Juyong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18633-18642
In this paper, we propose Neural Points, a novel point cloud representation and apply it to the arbitrary-factored upsampling task. Different from traditional point cloud representation where each point only represents a position or a local plane in the 3D space, each point in Neural Points represents a local continuous geometric shape via neural fields. Therefore, Neural Points contain more shape information and thus have a stronger representation ability. Neural Points is trained with surface containing rich geometric details, such that the trained model has enough expression ability for various shapes. Specifically, we extract deep local features on the points and construct neural fields through the local isomorphism between the 2D parametric domain and the 3D local patch. In the final, local neural fields are integrated together to form the global surface. Experimental results show that Neural Points has powerful representation ability and demonstrate excellent robustness and generalization ability. With Neural Points, we can resample point cloud with arbitrary resolutions, and it outperforms the state-of-the-art point cloud upsampling methods. Code is available at https://github.com/WanquanF/NeuralPoints.
****
Distribution Consistent Neural Architecture Search
Junyi Pan, Chong Sun, Yizhou Zhou, Ying Zhang, Chen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10884-10893
Recent progress on neural architecture search (NAS) has demonstrated exciting results on automating deep network architecture designs. In order to overcome the unaffordable complexity of training each candidate architecture from scratch, the state-of-the-art one-shot NAS approaches adopt a weight-sharing strategy to improve training efficiency. Although the computational cost is greatly reduced, such one-shot process introduces a severe weight coupling problem that largely degrades the evaluation accuracy of each candidate. The existing approaches often address the problem by shrinking the search space, model distillation, or few-shot training. Instead, in this paper, we propose a novel distribution consistent one-shot neural architecture search algorithm. We first theoretically investigate how the weight coupling problem affects the network searching performance from a parameter distribution perspective, and then propose a novel supernet training strategy with a Distribution Consistent Constraint that can provide a good measurement for the extent to which two architectures can share weights. Our strategy optimizes the supernet through iteratively inferring network weights and corresponding local sharing states. Such joint optimization of supernet's weights and topologies can diminish the discrepancy between the weights inherited from the supernet and the ones that are trained with a stand-alone model. As a result, it enables a more accurate model evaluation phase and leads to a better searching performance. We conduct extensive experiments on benchmark datasets with multiple searching spaces. The resulting architecture achieves superior performance over the current state-of-the-art NAS algorithms with comparable search costs, which demonstrates the efficacy of our approach.
****
Video-Text Representation Learning via Differentiable Weak Temporal Alignment
Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, Hyunwoo J. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5016-5025
Learning generic joint representations for video and text by a supervised method requires a prohibitively substantial amount of manually annotated video datasets. As a practical alternative, a large-scale but uncurated and narrated video dataset, HowTo100M, has recently been introduced. But it is still challenging to learn joint embeddings of video and text in a self-supervised manner, due to its ambiguity and non-sequential alignment. In this paper, we propose a novel multi-

modal self-supervised framework Video-Text Temporally Weak Alignment-based Contrastive Learning (VT-TWINS) to capture significant information from noisy and weakly correlated data using a variant of Dynamic Time Warping (DTW). We observe that the standard DTW inherently cannot handle weakly correlated data and only considers the globally optimal alignment path. To address these problems, we develop a differentiable DTW which also reflects local information with weak temporal alignment. Moreover, our proposed model applies a contrastive learning scheme to learn feature representations on weakly correlated data. Our extensive experiments demonstrate that VT-TWINS attains significant improvements in multi-modal representation learning and outperforms various challenging downstream tasks. Code is available at https://github.com/mlvlab/VT-TWINS.

****

Bi-Directional Object-Context Prioritization Learning for Saliency Ranking
Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5882-5891

The saliency ranking task is recently proposed to study the visual behavior that humans would typically shift their attention over different objects of a scene based on their degrees of saliency. Existing approaches focus on learning either object-object or object-scene relations. Such a strategy follows the idea of object-based attention in Psychology, but it tends to favor those objects with strong semantics (e.g., humans), resulting in unrealistic saliency ranking. We observe that spatial attention works concurrently with object-based attention in the human visual recognition system. During the recognition process, the human spatial attention mechanism would move, engage, and disengage from region to region (i.e., context to context). This inspires us to model the region-level interactions, in addition to the object-level reasoning, for saliency ranking. To this end, we propose a novel bi-directional method to unify spatial attention and object-based attention for saliency ranking. Our model includes two novel modules: (1) a selective object saliency (SOS) module that models object-based attention via inferring the semantic representation of the salient object, and (2) an object-context-object relation (OCOR) module that allocates saliency ranks to objects by jointly modeling the object-context and context-object interactions of the salient objects. Extensive experiments show that our approach outperforms existing state-of-the-art methods. Code and pretrained model are available at https://github.com/GrassBro/OCOR.

****

FreeSOLO: Learning To Segment Objects Without Annotations
Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, Jose M. Alvarez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14176-14186

Instance segmentation is a fundamental vision task that aims to recognize and segment each object in an image. However, it requires costly annotations such as bounding boxes and segmentation masks for learning. In this work, we propose a fully unsupervised learning method that learns class-agnostic instance segmentation without any annotations. We present FreeSOLO, a self-supervised instance segmentation framework built on top of the simple instance segmentation method SOLO. Our method also presents a novel localization-aware pre-training framework, where objects can be discovered from complicated scenes in an unsupervised manner. FreeSOLO achieves 9.8% AP50 on the challenging COCO dataset, which even outperforms several segmentation proposal methods that use manual annotations. For the first time, we demonstrate unsupervised class-agnostic instance segmentation successfully. FreeSOLO's box localization significantly outperforms state-of-the-art unsupervised object detection/discovery methods, with about 100% relative improvements in COCO AP. FreeSOLO further demonstrates superiority as a strong pre-training method, outperforming state-of-the-art self-supervised pre-training methods by +9.8% AP when fine-tuning instance segmentation with only 5% COCO masks.

****

What Do Navigation Agents Learn About Their Environment?
Kshitij Dwivedi, Gemma Roig, Aniruddha Kembhavi, Roozbeh Mottaghi; Proceedings o

f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
2, pp. 10276-10285
Today's state of the art visual navigation agents typically consist of large dee
p learning architectures trained end to end. Such models offer little to no inte
rpretability about the skills learned by the agent or the actions taken by it in
 response to its environment. While past works have explored interpreting deep l
earning models, little attention has been devoted to interpreting embodied AI sy
stems, which often involve reasoning about the structure of the environment, tar
get characteristics and the outcome of one's actions. In this paper, we introduc
e the Interpretability System for Embodied agEnts (iSEE) for Point Goal (PointNa
v) and Object Goal (ObjectNav) navigation models. We use iSEE to probe the dynam
ic representations produced by PointNav and ObjectNav agents for the presence of
 information about their agents location and actions, as well as the environment
. We demonstrate interesting insights about navigation agents using iSEE, includ
ing the ability to encode reachable locations (to avoid obstacles), visibility o
f the target, progress from the initial spawn location as well as the dramatic e
ffect on the behaviors of agents when we mask out critical individual neurons.
****

Progressive Minimal Path Method With Embedded CNN
Wei Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2022, pp. 4514-4522
We propose Path-CNN, a method for the segmentation of centerlines of tubular str
uctures by embedding convolutional neural networks (CNNs) into the progressive m
inimal path method. Minimal path methods are widely used for topology-aware cent
erline segmentation, but usually these methods rely on weak, hand-tuned image fe
atures. In contrast, CNNs use strong image features which are learned automatica
lly from images. But CNNs usually do not take the topology of the results into a
ccount, and often require a large amount of annotations for training. We integra
te CNNs into the minimal path method, so that both techniques benefit from each
other: CNNs employ learned image features to improve the determination of minima
l paths, while the minimal path method ensures the correct topology of the segme
nted centerlines, provides strong geometric priors to increase the performance o
f CNNs, and reduces the amount of annotations for the training of CNNs significa
ntly. Our method has lower hardware requirements than many recent methods. Quali
tative and quantitative comparison with other methods shows that Path-CNN achiev
es better performance, especially when dealing with tubular structures with comp
lex shapes in challenging environments.
****

FIFO: Learning Fog-Invariant Features for Foggy Scene Segmentation
Sohyun Lee, Taeyoung Son, Suha Kwak; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2022, pp. 18911-18921
Robust visual recognition under adverse weather conditions is of great importanc
e in real-world applications. In this context, we propose a new method for learn
ing semantic segmentation models robust against fog. Its key idea is to consider
 the fog condition of an image as its style and close the gap between images wit
h different fog conditions in neural style spaces of a segmentation model. In pa
rticular, since the neural style of an image is in general affected by other fac
tors as well as fog, we introduce a fog-pass filter module that learns to extrac
t a fog-relevant factor from the style. Optimizing the fog-pass filter and the s
egmentation model alternately gradually closes the style gap between different f
og conditions and allows to learn fog-invariant features in consequence. Our met
hod substantially outperforms previous work on three real foggy image datasets.
Moreover, it improves performance on both foggy and clear weather images, while
existing methods often degrade performance on clear scenes.
****

3D Human Tongue Reconstruction From Single "In-the-Wild" Images
Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, Stefanos Zafe
iriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2022, pp. 2771-2780
3D face reconstruction from a single image is a task that has garnered increased

interest in the Computer Vision community, especially due to its broad use in a number of applications such as realistic 3D avatar creation, pose invariant face recognition and face hallucination. Since the introduction of the 3D Morphable Model in the late 90's, we witnessed an explosion of research aiming at particularly tackling this task. Nevertheless, despite the increasing level of detail in the 3D face reconstructions from single images mainly attributed to deep learning advances, finer and highly deformable components of the face such as the tongue are still absent from all 3D face models in the literature, although being very important for the realness of the 3D avatar representations. In this work we present the first, to the best of our knowledge, end-to-end trainable pipeline that accurately reconstructs the 3D face together with the tongue. Moreover, we make this pipeline robust in "in-the-wild" images by introducing a novel GAN method tailored for 3D tongue surface generation. Finally, we make publicly available to the community the first diverse tongue dataset, consisting of 1,800 raw scans of 700 individuals varying in gender, age, and ethnicity backgrounds. As we demonstrate in an extensive series of quantitative as well as qualitative experiments, our model proves to be robust and realistically captures the 3D tongue structure, even in adverse "in-the-wild" conditions.

****

Enhancing Adversarial Robustness for Deep Metric Learning

Mo Zhou, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15325-15334

Owing to security implications of adversarial vulnerability, adversarial robustness of deep metric learning models has to be improved. In order to avoid model collapse due to excessively hard examples, the existing defenses dismiss the min-max adversarial training, but instead learn from a weak adversary inefficiently. Conversely, we propose Hardness Manipulation to efficiently perturb the training triplet till a specified level of hardness for adversarial training, according to a harder benign triplet or a pseudo-hardness function. It is flexible since regular training and min-max adversarial training are its boundary cases. Besides, Gradual Adversary, a family of pseudo-hardness functions is proposed to gradually increase the specified hardness level during training for a better balance between performance and robustness. Additionally, an Intra-Class Structure loss term among benign and adversarial examples further improves model robustness and efficiency. Comprehensive experimental results suggest that the proposed method, although simple in its form, overwhelmingly outperforms the state-of-the-art defenses in terms of robustness, training efficiency, as well as performance on benign examples.

****

Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation

Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, David Z. Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12094-12103

Vision Transformers (ViTs) have emerged with superior performance on computer vision tasks compared to convolutional neural network (CNN)-based models. However, ViTs are mainly designed for image classification that generate single-scale low-resolution representations, which makes dense prediction tasks such as semantic segmentation challenging for ViTs. Therefore, we propose HRViT, which enhances ViTs to learn semantically-rich and spatially-precise multi-scale representations by integrating high-resolution multi-branch architectures with ViTs. We balance the model performance and efficiency of HRViT by various branch-block co-optimization techniques. Specifically, we explore heterogeneous branch designs, reduce the redundancy in linear layers, and augment the attention block with enhanced expressiveness. Those approaches enabled \ours to push the Pareto frontier of performance and efficiency on semantic segmentation to a new level, as our evaluation results on ADE20K and Cityscapes show. HRViT achieves 50.20% mIoU on ADE20K and 83.16% mIoU on Cityscapes for semantic segmentation tasks, surpassing state-of-the-art MiT and CSWin backbones with an average of +1.78 mIoU improvement, 28% parameter reduction, and 21% FLOPs reduction, demonstrating the potential of HRViT as a strong vision backbone for semantic segmentation.

****
Lite-MDETR: A Lightweight Multi-Modal Detector
Qian Lou, Yen-Chang Hsu, Burak Uzkent, Ting Hua, Yilin Shen, Hongxia Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12206-12215
Recent multi-modal detectors based on transformers and modality encoders have successfully achieved impressive results on end-to-end visual object detection conditioned on a raw text query. However, they require a large model size and an enormous amount of computations to achieve high performance, which makes it difficult to deploy mobile applications that are limited by tight hardware resources. In this paper, we present a Lightweight modulated detector, Lite-MDETR, to facilitate efficient end-to-end multi-modal understanding on mobile devices. The key primitive is that Dictionary-Lookup-Transformormations (DLT) is proposed to replace Linear Transformation (LT) in multi-modal detectors where each weight in Linear Transformation (LT) is approximately factorized into a smaller dictionary, index, and coefficient. This way, the enormous linear projection with weights is converted into lite linear projection with dictionaries, a few lookups and scalings with indices and coefficients. DLT can be directly applied to pre-trained detectors, removing the need to perform expensive training from scratch. To tackle the challenging training of DLT due to the non-differentiable index, we convert the index and coefficient into a sparse matrix, train this sparse matrix during the fine-tuning phase, and recover it back to index and coefficient during the inference phase. Extensive experiments on several tasks such as phrase grounding, referring expression comprehension and segmentation show that our Lite-MDETR achieves similar detection accuracy to the prior multi-modal detectors with ~ 4.1 xmodel size reduction.
****
CoordGAN: Self-Supervised Dense Correspondences Emerge From GANs
Jiteng Mu, Shalini De Mello, Zhiding Yu, Nuno Vasconcelos, Xiaolong Wang, Jan Kautz, Sifei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10011-10020
Recent advances show that Generative Adversarial Networks (GANs) can synthesize images with smooth variations along semantically meaningful latent directions, such as pose, expression, layout, etc. While this indicates that GANs implicitly learn pixel-level correspondences across images, few studies explored how to extract them explicitly. In this work, we introduce Coordinate GAN (CoordGAN), a structure-texture disentangled GAN that learns a dense correspondence map for each generated image. We represent the correspondence maps of different images as warped coordinate frames transformed from a canonical coordinate frame, i.e., the correspondence map, which describes the structure (e.g., the shape of a face), is controlled via a transformation. Hence, finding correspondences boils down to locating the same coordinate in different correspondence maps. In CoordGAN, we sample a transformation to represent the structure of a synthesized instance, while an independent texture branch is responsible for rendering appearance details orthogonal to the structure. Our approach can also extract dense correspondence maps for real images by adding an encoder on top of the generator. We quantitatively demonstrate the quality of the learned dense correspondences through segmentation mask transfer on multiple datasets. We also show that the proposed generator achieves better structure and texture disentanglement compared to existing approaches.
****
A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation
Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, Stephen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5120-5130
This paper proposes a simple transfer learning baseline for sign language translation. Existing sign language datasets (e.g. PHOENIX-2014T, CSL-Daily) contain only about 10K-20K pairs of sign videos, gloss annotations and texts, which are an order of magnitude smaller than typical parallel data for training spoken language translation models. Data is thus a bottleneck for training effective sign l

anguage translation models. To mitigate this problem, we propose to progressivel y pretrain the model from general-domain datasets that include a large amount of external supervision to within-domain datasets. Concretely, we pretrain the sig n-to-gloss visual network on the general domain of human actions and the within- domain of a sign-to-gloss dataset, and pretrain the gloss-to-text translation ne twork on the general domain of a multilingual corpus and the within-domain of a gloss-to-text corpus. The joint model is fine-tuned with an additional module na med the visual-language mapper that connects the two networks. This simple basel ine surpasses the previous state-of-the-art results on two sign language transla tion benchmarks, demonstrating the effectiveness of transfer learning. With its simplicity and strong performance, this approach can serve as a solid baseline f or future research.

****

Unsupervised Visual Representation Learning by Online Constrained K-Means

Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, Rong Jin; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16640-1664 9

Cluster discrimination is an effective pretext task for unsupervised representat ion learning, which often consists of two phases: clustering and discrimination. Clustering is to assign each instance a pseudo label that will be used to learn representations in discrimination. The main challenge resides in clustering sin ce prevalent clustering methods (e.g., k-means) have to run in a batch mode. Bes ides, there can be a trivial solution consisting of a dominating cluster. To add ress these challenges, we first investigate the objective of clustering-based re presentation learning. Based on this, we propose a novel clustering-based pretex t task with online Constrained K-means (CoKe). Compared with the balanced cluste ring that each cluster has exactly the same size, we only constrain the minimal size of each cluster to flexibly capture the inherent data structure. More impor tantly, our online assignment method has a theoretical guarantee to approach the global optimum. By decoupling clustering and discrimination, CoKe can achieve c ompetitive performance when optimizing with only a single view from each instanc e. Extensive experiments on ImageNet and other benchmark data sets verify both t he efficacy and efficiency of our proposal.

****

Neural Point Light Fields

Julian Ost, Issam Laradji, Alejandro Newell, Yuval Bahat, Felix Heide; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18419-18429

We introduce Neural Point Light Fields that represent scenes implicitly with a l ight field living on a sparse point cloud. Combining differentiable volume rende ring with learned implicit density representations has made it possible to synth esize photo-realistic images for novel views of small scenes. As neural volumetr ic rendering methods require dense sampling of the underlying functional scene r epresentation, at hundreds of samples along a ray cast through the volume, they are fundamentally limited to small scenes with the same objects projected to hun dreds of training views. Promoting sparse point clouds to neural implicit light fields allows us to represent large scenes effectively with only a single radian ce evaluation per ray. These point light fields are as a function of the ray dir ection, and local point feature neighborhood, allowing us to interpolate the lig ht field conditioned training images without dense object coverage and parallax. We assess the proposed method for novel view synthesis on large driving scenari os, where we synthesize realistic unseen views that existing implicit approaches fail to represent. We validate that Neural Point Light Fields make it possible to predict videos along unseen trajectories previously only feasible to generate by explicitly modeling the scene.

****

Vehicle Trajectory Prediction Works, but Not Everywhere

Mohammadhossein Bahari, Saeed Saadatnejad, Ahmad Rahimi, Mohammad Shaverdikondor i, Amir Hossein Shahidzadeh, Seyed-Mohsen Moosavi-Dezfooli, Alexandre Alahi; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (

CVPR), 2022, pp. 17123-17133
Vehicle trajectory prediction is nowadays a fundamental pillar of self-driving cars. Both the industry and research communities have acknowledged the need for such a pillar by providing public benchmarks. While state-of-the-art methods are impressive, i.e., they have no off-road prediction, their generalization to cities outside of the benchmark remains unexplored. In this work, we show that those methods do not generalize to new scenes. We present a novel method that automatically generates realistic scenes causing state-of-the-art models to go off-road. We frame the problem through the lens of adversarial scene generation. The method is a simple yet effective generative model based on atomic scene generation functions along with physical constraints. Our experiments show that more than 60% of existing scenes from the current benchmarks can be modified in a way to make prediction methods fail (i.e., predicting off-road). We further show that the generated scenes (i) are realistic since they do exist in the real world, and (ii) can be used to make existing models more robust, yielding 30-40% reductions in the off-road rate. The code is available online: https://s-attack.github.io/
****

PSMNet: Position-Aware Stereo Merging Network for Room Layout Estimation
Haiyan Wang, Will Hutchcroft, Yuguang Li, Zhiqiang Wan, Ivaylo Boyadzhiev, Yingli Tian, Sing Bing Kang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8616-8625
In this paper, we propose a new deep learning-based method for estimating room layout given a pair of 360 panoramas. Our system, called Position-aware Stereo Merging Network or PSMNet, is an end-to-end joint layout-pose estimator. PSMNet consists of a Stereo Pano Pose (SP^2) transformer and a novel Cross-Perspective Projection (CP^2) layer. The stereo-view SP^2 transformer is used to implicitly infer correspondences between views, and can handle noisy poses. The pose-aware CP^2layer is designed to render features from the adjacent view to the anchor (reference) view, in order to perform view fusion and estimate the visible layout. Our experiments and analysis validate our method, which significantly outperforms the state-of-the-art layout estimators, especially for large and complex room spaces.
****

MonoDTR: Monocular 3D Object Detection With Depth-Aware Transformer
Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, Winston H. Hsu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4012-4021
Monocular 3D object detection is an important yet challenging task in autonomous driving. Some existing methods leverage depth information from an off-the-shelf depth estimator to assist 3D detection, but suffer from the additional computational burden and achieve limited performance caused by inaccurate depth priors. To alleviate this, we propose MonoDTR, a novel end-to-end depth-aware transformer network for monocular 3D object detection. It mainly consists of two components: (1) the Depth-Aware Feature Enhancement (DFE) module that implicitly learns depth-aware features with auxiliary supervision without requiring extra computation, and (2) the Depth-Aware Transformer (DTR) module that globally integrates context- and depth-aware features. Moreover, different from conventional pixel-wise positional encodings, we introduce a novel depth positional encoding (DPE) to inject depth positional hints into transformers. Our proposed depth-aware modules can be easily plugged into existing image-only monocular 3D object detectors to improve the performance. Extensive experiments on the KITTI dataset demonstrate that our approach outperforms previous state-of-the-art monocular-based methods and achieves real-time detection. Code is available at https://github.com/kuanchihhuang/MonoDTR.
****

Learning Graph Regularisation for Guided Super-Resolution
Riccardo de Lutio, Alexander Becker, Stefano D'Aronco, Stefania Russo, Jan D. Wegner, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1979-1988
We introduce a novel formulation for guided super-resolution. Its core is a diff

erentiable optimisation layer that operates on a learned affinity graph. The lea rned graph potentials make it possible to leverage rich contextual information f rom the guide image, while the explicit graph optimisation within the architectu re guarantees rigorous fidelity of the high-resolution target to the low-resolut ion source. With the decision to employ the source as a constraint rather than o nly as an input to the prediction, our method differs from state-of-the-art deep architectures for guided super-resolution, which produce targets that, when dow nsampled, will only approximately reproduce the source. This is not only theoret ically appealing, but also produces crisper, more natural-looking images. A key property of our method is that, although the graph connectivity is restricted to the pixel lattice, the associated edge potentials are learned with a deep featu re extractor and can encode rich context information over large receptive fields . By taking advantage of the sparse graph connectivity, it becomes possible to p ropagate gradients through the optimisation layer and learn the edge potentials from data. We extensively evaluate our method on several datasets, and consisten tly outperform recent baselines in terms of quantitative reconstruction errors, while also delivering visually sharper outputs. Moreover, we demonstrate that ou r method generalises particularly well to new datasets not seen during training.
****

Instance-Wise Occlusion and Depth Orders in Natural Scenes
Hyunmin Lee, Jaesik Park; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2022, pp. 21210-21221
In this paper, we introduce a new dataset, named InstaOrder, that can be used to understand the spatial relationships of instances in a 3D space. The dataset co nsists of 2.9M annotations of geometric orderings for class-labeled instances in 101K natural scenes. The scenes were annotated by 3,659 crowd-workers regarding (1) occlusion order that identifies occluder/occludee and (2) depth order that describes ordinal relations that consider relative distance from the camera. The dataset provides joint annotation of two kinds of orderings for the same instan ces, and we discover that the occlusion order and depth order are complementary. We also introduce a geometric order prediction network called InstaOrderNet, wh ich is superior to state-of-the-art approaches. Moreover, we propose a dense dep th prediction network called InstaDepthNet that uses auxiliary geometric order l oss to boost the accuracy of the state-of-the-art depth prediction approach, MiD aS [54].
****

Look for the Change: Learning Object States and State-Modifying Actions From Unt rimmed Web Videos
Tomáš Sou■ek, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, Josef Sivic; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13956-13966
Human actions often induce changes of object states such as "cutting an apple", "cleaning shoes" or "pouring coffee". In this paper, we seek to temporally local ize object states (e.g. "empty" and "full" cup) together with the corresponding state-modifying actions ("pouring coffee") in long uncurated videos with minimal supervision. The contributions of this work are threefold. First, we develop a self-supervised model for jointly learning state-modifying actions together with the corresponding object states from an uncurated set of videos from the Intern et. The model is self-supervised by the causal ordering signal, i.e. initial obj ect state -> manipulating action -> end state. Second, to cope with noisy uncura ted training data, our model incorporates a noise adaptive weighting module supe rvised by a small number of annotated still images, that allows to efficiently f ilter out irrelevant videos during training. Third, we collect a new dataset wit h more than 2600 hours of video and 34 thousand changes of object states, and ma nually annotate a part of this data to validate our approach. Our results demons trate substantial improvements over prior work in both action and object state-r ecognition in video.
****

Attribute Surrogates Learning and Spectral Tokens Pooling in Transformers for Fe w-Shot Learning

Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, Wen qiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 9119-9129

This paper presents new hierarchically cascaded transformers that can improve da ta efficiency through attribute surrogates learning and spectral tokens pooling. Vision transformers have recently been thought of as a promising alternative to convolutional neural networks for visual recognition. But when there is no suff icient data, it gets stuck in overfitting and shows inferior performance. To imp rove data efficiency, we propose hierarchically cascaded transformers that explo it intrinsic image structures through spectral tokens pooling and optimize the l earnable parameters through latent attribute surrogates. The intrinsic image str ucture is utilized to reduce the ambiguity between foreground content and backgr ound noise by spectral tokens pooling. And the attribute surrogate learning sche me is designed to benefit from the rich visual information in image-label pairs instead of simple visual concepts assigned by their labels. Our Hierarchically C ascaded Transformers, called HCTransformers, is built upon a self-supervised lea rning framework DINO and is tested on several popular few-shot learning benchmar ks. In the inductive setting, HCTransformers surpass the DINO baseline by a larg e margin of 9.7% 5-way 1-shot accuracy and 9.17% 5-way 5-shot accuracy on mini-I mageNet, which demonstrates HCTransformers are efficient to extract discriminati ve features. Also, HCTransformers show clear advantages over SOTA few-shot class ification methods in both 5-way 1-shot and 5-way 5-shot settings on four popular benchmark datasets, including mini-ImageNet, tiered-ImageNet, FC100, and CIFAR- FS. The trained weights and codes are available at https://github.com/StomachCol d/HCTransformers.

****

Generalized Category Discovery

Sagar Vaze, Kai Han, Andrea Vedaldi, Andrew Zisserman; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7492- 7501

In this paper, we consider a highly general image recognition setting wherein, g iven a labelled and unlabelled set of images, the task is to categorize all imag es in the unlabelled set. Here, the unlabelled images may come from labelled cla sses or from novel ones. Existing recognition methods are not able to deal with this setting, because they make several restrictive assumptions, such as the unl abelled instances only coming from known -- or unknown -- classes, and the numbe r of unknown classes being known a-priori. We address the more unconstrained set ting, naming it 'Generalized Category Discovery', and challenge all these assump tions. We first establish strong baselines by taking state-of-the-art algorithms from novel category discovery and adapting them for this task. Next, we propose the use of vision transformers with contrastive representation learning for thi s open-world setting. We then introduce a simple yet effective semi-supervised k -means method to cluster the unlabelled data into seen and unseen classes automa tically, substantially outperforming the baselines. Finally, we also propose a n ew approach to estimate the number of classes in the unlabelled data. We thoroug hly evaluate our approach on public datasets for generic object classification a nd on fine-grained datasets, leveraging the recent Semantic Shift Benchmark suit e. Code: https://www.robots.ox.ac.uk/~vgg/research/gcd

****

Maximum Consensus by Weighted Influences of Monotone Boolean Functions

Erchuan Zhang, David Suter, Ruwan Tennakoon, Tat-Jun Chin, Alireza Bab-Hadiashar , Giang Truong, Syed Zulqarnain Gilani; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8964-8972

Maximisation of Consensus (MaxCon) is one of the most widely used robust criteri a in computer vision. Tennakoon et al. (CVPR2021), made a connection between Max Con and estimation of influences of a Monotone Boolean function. In such, there are two distributions involved: the distribution defining the influence measure; and the distribution used for sampling to estimate the influence measure. This paper studies the concept of weighted influences for solving MaxCon. In particul ar, we study the Bernoulli measures. Theoretically, we prove the weighted influe

nces, under this measure, of points belonging to larger structures are smaller than those of points belonging to smaller structures in general. We also consider another "natural" family of weighting strategies: sampling with uniform measure concentrated on a particular (Hamming) level of the cube. One can choose to have matching distributions: the same for defining the measure as for implementing the sampling. This has the advantage that the sampler is an unbiased estimator of the measure. Based on weighted sampling, we modify the algorithm of Tennakoon et al., and test on both synthetic and real datasets. We show some modest gains of Bernoulli sampling, and we illuminate some of the interactions between structure in data and weighted measures and weighted sampling.
****

TransforMatcher: Match-to-Match Attention for Semantic Correspondence
Seungwook Kim, Juhong Min, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8697-8707
Establishing correspondences between images remains a challenging task, especially under large appearance changes due to different viewpoints or intra-class variations. In this work, we introduce a strong semantic image matching learner, dubbed TransforMatcher, which builds on the success of transformer networks in vision domains. Unlike existing convolution- or attention-based schemes for correspondence, TransforMatcher performs global match-to-match attention for precise match localization and dynamic refinement. To handle a large number of matches in a dense correlation map, we develop a light-weight attention architecture to consider the global match-to-match interactions. We also propose to utilize a multi-channel correlation map for refinement, treating the multi-level scores as features instead of a single score to fully exploit the richer layer-wise semantics. In experiments, TransforMatcher sets a new state of the art on SPair-71k while performing on par with existing SOTA methods on the PF-PASCAL dataset.
****

Robust Outlier Detection by De-Biasing VAE Likelihoods
Kushal Chauhan, Barath Mohan U, Pradeep Shenoy, Manish Gupta, Devarajan Sridharan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9881-9890
Deep networks often make confident, yet, incorrect, predictions when tested with outlier data that is far removed from their training distributions. Likelihoods computed by deep generative models (DGMs) are a candidate metric for outlier detection with unlabeled data. Yet, previous studies have shown that DGM likelihoods are unreliable and can be easily biased by simple transformations to input data. Here, we examine outlier detection with variational autoencoders (VAEs), among the simplest of DGMs. We propose novel analytical and algorithmic approaches to ameliorate key biases with VAE likelihoods. Our bias corrections are sample-specific, computationally inexpensive, and readily computed for various decoder visible distributions. Next, we show that a well-known image pre-processing technique -- contrast stretching -- extends the effectiveness of bias correction to further improve outlier detection. Our approach achieves state-of-the-art accuracies with nine grayscale and natural image datasets, and demonstrates significant advantages -- both with speed and performance -- over four recent, competing approaches. In summary, lightweight remedies suffice to achieve robust outlier detection with VAEs.
****

Contour-Hugging Heatmaps for Landmark Detection
James McCouat, Irina Voiculescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20597-20605
We propose an effective and easy-to-implement method for simultaneously performing landmark detection in images and obtaining an ingenious uncertainty measurement for each landmark. Uncertainty measurements for landmarks are particularly useful in medical imaging applications: rather than giving an erroneous reading, a landmark detection system is more useful when it flags its level of confidence in its prediction. When an automated system is unsure of its predictions, the accuracy of the results can be further improved manually by a human. In the medical domain, being able to review an automated system's level of certainty signific

antly improves a clinician's trust in it. This paper obtains landmark predictions with uncertainty measurements using a three stage method: 1) We train our network on one-hot heatmap images, 2) We calibrate the uncertainty of the network using temperature scaling, 3) We calculate a novel statistic called 'Expected Radial Error' to obtain uncertainty measurements. We find that this method not only achieves localisation results on par with other state-of-the-art methods but also an uncertainty score which correlates with the true error for each landmark thereby bringing an overall step change in what a generic computer vision method for landmark detection should be capable of. In addition, we show that our uncertainty measurement can be used to classify, with good accuracy, what landmark predictions are likely to be inaccurate. Code available at: https://github.com/jfm15/ContourHuggingHeatmaps.git

****

Voxel Field Fusion for 3D Object Detection

Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1120-1129

In this work, we present a conceptually simple yet effective framework for cross-modality 3D object detection, named voxel field fusion. The proposed approach aims to maintain cross-modality consistency by representing and fusing augmented image features as a ray in the voxel field. To this end, the learnable sampler is first designed to sample vital features from the image plane that are projected to the voxel grid in a point-to-ray manner, which maintains the consistency in feature representation with spatial context. In addition, ray-wise fusion is conducted to fuse features with the supplemental context in the constructed voxel field. We further develop mixed augmentor to align feature-variant transformations, which bridges the modality gap in data augmentation. The proposed framework is demonstrated to achieve consistent gains in various benchmarks and outperforms previous fusion-based methods on KITTI and nuScenes datasets. Code is made available at https://github.com/dvlab-research/VFF.

****

Divide and Conquer: Compositional Experts for Generalized Novel Class Discovery

Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, Cheng Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14268-14277

In response to the explosively-increasing requirement of annotated data, Novel Class Discovery (NCD) has emerged as a promising alternative to automatically recognize unknown classes without any annotation. To this end, a model makes use of a base set to learn basic semantic discriminability that can be transferred to recognize novel classes. Most existing works handle the base and novel sets using separate objectives within a two-stage training paradigm. Despite showing competitive performance on novel classes, they fail to generalize to recognizing samples from both base and novel sets. In this paper, we focus on this generalized setting of NCD (GNCD), and propose to divide and conquer it with two groups of Compositional Experts (ComEx). Each group of experts is designed to characterize the whole dataset in a comprehensive yet complementary fashion. With their union, we can solve GNCD in an efficient end-to-end manner. We further look into the drawback in current NCD methods, and propose to strengthen ComEx with global-to-local and local-to-local regularization. ComEx is evaluated on four popular benchmarks, showing clear superiority towards the goal of GNCD.

****

Programmatic Concept Learning for Human Motion Description and Synthesis

Sumith Kulal, Jiayuan Mao, Alex Aiken, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13843-13852

We introduce Programmatic Motion Concepts, a hierarchical motion representation for human actions that captures both low level motion and high level description as motion concepts. This representation enables human motion description, interactive editing, and controlled synthesis of novel video sequences within a single framework. We present an architecture that learns this concept representation

from paired video and action sequences in a semi-supervised manner. The compactn ess of our representation also allows us to present a low-resource training reci pe for data-efficient learning. By outperforming established baselines, especial ly in small data regime, we demonstrate the efficiency and effectiveness of our framework for multiple applications.
****

Interpretable Part-Whole Hierarchies and Conceptual-Semantic Relationships in Ne ural Networks
Nicola Garau, Niccolò Bisagno, Zeno Sambugaro, Nicola Conci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13689-13698
Deep neural networks achieve outstanding results in a large variety of tasks, of ten outperforming human experts. However, a known limitation of current neural a rchitectures is the poor accessibility to understand and interpret the network r esponse to a given input. This is directly related to the huge number of variabl es and the associated non-linearities of neural models, which are often used as black boxes. When it comes to critical applications as autonomous driving, secur ity and safety, medicine and health, the lack of interpretability of the network behavior tends to induce skepticism and limited trustworthiness, despite the ac curate performance of such systems in the given task. Furthermore, a single metr ic, such as the classification accuracy, provides a non-exhaustive evaluation of most real-world scenarios. In this paper, we want to make a step forward toward s interpretability in neural networks, providing new tools to interpret their be havior. We present Agglomerator, a framework capable of providing a representati on of part-whole hierarchies from visual cues and organizing the input distribut ion matching the conceptual-semantic hierarchical structure between classes. We evaluate our method on common datasets, such as SmallNORB, MNIST, FashionMNIST, CIFAR-10, and CIFAR-100, providing a more interpretable model than other state-o f-the-art approaches.
****

Fast Algorithm for Low-Rank Tensor Completion in Delay-Embedded Space
Ryuki Yamamoto, Hidekata Hontani, Akira Imakura, Tatsuya Yokota; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2058-2066
Tensor completion using multiway delay-embedding transform (MDT) (or Hankelizati on) suffers from the large memory requirement and high computational cost in spi te of its high potentiality for the image modeling. Recent studies have shown hi gh completion performance with a relatively small window size, but experiments w ith large window sizes require huge amount of memory and cannot be easily calcul ated. In this study, we address this serious computational issue, and propose it s fast and efficient algorithm. Key techniques of the proposed method are based on two properties: (1) the signal after MDT can be diagonalized by Fourier trans form, (2) an inverse MDT can be represented as a convolutional form. To use the properties, we modify MDT-Tucker, a method using Tucker decomposition with MDT, and introducing the fast and efficient algorithm. Our experiments show more than 100 times acceleration while maintaining high accuracy, and to realize the comp utation with large window size.
****

Panoptic, Instance and Semantic Relations: A Relational Context Encoder To Enhan ce Panoptic Segmentation
Shubhankar Borse, Hyojin Park, Hong Cai, Debasmit Das, Risheek Garrepalli, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1269-1279
This paper presents a novel framework to integrate both semantic and instance co ntexts for panoptic segmentation. In existing works, it is common to use a share d backbone to extract features for both things (countable classes such as vehicl es) and stuff (uncountable classes such as roads). This, however, fails to captu re the rich relations among them, which can be utilized to enhance visual unders tanding and segmentation performance. To address this shortcoming, we propose a novel Panoptic, Instance, and Semantic Relations (PISR) module to exploit such c

ontexts. First, we generate panoptic encodings to summarize key features of the semantic classes and predicted instances. A Panoptic Relational Attention (PRA) module is then applied to the encodings and the global feature map from the back bone. It produces a feature map that captures 1) the relations across semantic c lasses and instances and 2) the relations between these panoptic categories and spatial features. PISR also automatically learns to focus on the more important instances, making it robust to the number of instances used in the relational at tention module. Moreover, PISR is a general module that can be applied to any ex isting panoptic segmentation architecture. Through extensive evaluations on pano ptic segmentation benchmarks like Cityscapes, COCO, and ADE20K, we show that PIS R attains considerable improvements over existing approaches.
****

Point2Seq: Detecting 3D Objects As Sequences
Yujing Xue, Jiageng Mao, Minzhe Niu, Hang Xu, Michael Bi Mi, Wei Zhang, Xiaogang Wang, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2022, pp. 8521-8530
We present a simple and effective framework, named Point2Seq, for 3D object dete ction from point clouds. In contrast to previous methods that normally predict a ttributes of 3D objects all at once, we expressively model the interdependencies between attributes of 3D objects, which in turn enables a better detection accu racy. Specifically, we view each 3D object as a sequence of words and reformulat e the 3D object detection task as decoding words from 3D scenes in an auto-regre ssive manner. We further propose a lightweight scene-to-sequence decoder that ca n auto-regressively generate words conditioned on features from a 3D scene as we ll as cues from the preceding words. The predicted words eventually constitute a set of sequences that completely describe the 3D objects in the scene, and all the predicted sequences are then automatically assigned to the respective ground truths through similarity-based sequence matching. Our approach is conceptually intuitive and can be readily plugged upon most existing 3D-detection backbones without adding too much computational overhead; the sequential decoding paradigm we proposed, on the other hand, can better exploit information from complex 3D scenes with the aid of preceding predicted words. Without bells and whistles, ou r method significantly outperforms the previous anchor- and center-based 3D obje ct detection frameworks, yielding the new state-of-the-art on the challenging ON CE dataset as well as the Waymo Open Dataset. Code will be made publicly availab le.
****

Less Is More: Generating Grounded Navigation Instructions From Landmarks
Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Iz zeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, Peter Anderson; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 15428-15438
We study the automatic generation of navigation instructions from 360-degree ima ges captured on indoor routes. Existing generators suffer from poor visual groun ding, causing them to rely on language priors and hallucinate objects. Our MARKY -MT5 system addresses this by focusing on visual landmarks; it comprises a first stage landmark detector and a second stage generator--a multimodal, multilingua l, multitask encoder-decoder. To train it, we bootstrap grounded landmark annota tions on top of the Room-across-Room (RxR) dataset. Using text parsers, weak sup ervision from RxR's pose traces, and a multilingual image-text encoder trained o n 1.8b images, we identify 1.1m English, Hindi and Telugu landmark descriptions and ground them to specific regions in panoramas. On Room-to-Room, human wayfind ers obtain success rates (SR) of 73% following MARKY-MT5's instructions, just sh y of their 76% SR following human instructions---and well above SRs with other g enerators. Evaluations on RxR's longer, diverse paths obtain 62-64% SRs on three languages. Generating such high-quality navigation instructions in novel enviro nments is a step towards conversational navigation tools and could facilitate la rger-scale training of instruction-following agents.
****

Task-Adaptive Negative Envision for Few-Shot Open-Set Recognition

Shiyuan Huang, Jiawei Ma, Guangxing Han, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7171-7180

We study the problem of few-shot open-set recognition (FSOR), which learns a recognition system capable of both fast adaptation to new classes with limited labeled examples and rejection of unknown negative samples. Traditional large-scale open-set methods have been shown ineffective for FSOR problem due to data limitation. Current FSOR methods typically calibrate few-shot closed-set classifiers to be sensitive to negative samples so that they can be rejected via thresholding. However, threshold tuning is a challenging process as different FSOR tasks may require different rejection powers. In this paper, we instead propose task-adaptive negative class envision for FSOR to integrate threshold tuning into the learning process. Specifically, we augment the few-shot closed-set classifier with additional negative prototypes generated from few-shot examples. By incorporating few-shot class correlations in the negative generation process, we are able to learn dynamic rejection boundaries for FSOR tasks. Besides, we extend our method to generalized few-shot open-set recognition (GFSOR), which requires classification on both many-shot and few-shot classes as well as rejection of negative samples. Extensive experiments on public benchmarks validate our methods on both problems.

****

DisARM: Displacement Aware Relation Module for 3D Detection

Yao Duan, Chenyang Zhu, Yuqing Lan, Renjiao Yi, Xinwang Liu, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16980-16989

We introduce Displacement Aware Relation Module (DisARM), a novel neural network module for enhancing the performance of 3D object detection in point cloud scenes. The core idea is extracting the most principal contextual information is critical for detection while the target is incomplete or featureless. We find that relations between proposals provide a good representation to describe the context. However, adopting relations between all the object or patch proposals for detection is inefficient, and an imbalanced combination of local and global relations brings extra noise that could mislead the training. Rather than working with all relations, we find that training with relations only between the most representative ones, or anchors, can significantly boost the detection performance. Good anchors should be semantic-aware with no ambiguity and able to describe the whole layout of a scene with no redundancy. To find the anchors, we first perform a preliminary relation anchor module with an objectness-aware sampling approach and then devise a displacement based module for weighing the relation importance for better utilization of contextual information. This light-weight relation module leads to significantly higher accuracy of object instance detection when being plugged into the state-of- the-art detectors. Evaluations on the public benchmarks of real-world scenes show that our method achieves the state-of-the-art performance on both SUN RGB-D and ScanNet V2. The code and models are publicly available at https://github.com/YaraDuan/DisARM.

****

ETHSeg: An Amodel Instance Segmentation Network and a Real-World Dataset for X-Ray Waste Inspection

Lingteng Qiu, Zhangyang Xiong, Xuhao Wang, Kenkun Liu, Yihan Li, Guanying Chen, Xiaoguang Han, Shuguang Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2283-2292

Waste inspection for packaged waste is an important step in the pipeline of waste disposal. Previous methods either rely on manual visual checking or RGB image-based inspection algorithm, requiring costly preparation procedures (e.g., open the bag and spread the waste items). Moreover, occluded items are very likely to be left out. Inspired by the fact that X-ray has a strong penetrating power to see through the bag and overlapping objects, we propose to perform waste inspection efficiently using X-ray images without the need to open the bag. We introduce a novel problem of instance-level waste segmentation in X-ray image for intelligent waste inspection, and contribute a real dataset consisting of 5,038 X-ray

images (totally 30,881 waste items) with high-quality annotations (i.e., waste categories, object boxes, and instance-level masks) as a benchmark for this problem. As existing segmentation methods are mainly designed for natural images and cannot take advantage of the characteristics of X-ray waste images (e.g., heavy occlusions and penetration effect), we propose a new instance segmentation method to explicitly take these image characteristics into account. Specifically, our method adopts an easy-to-hard disassembling strategy to use high confidence predictions to guide the segmentation of highly overlapped objects, and a global structure guidance module to better capture the complex contour information caused by the penetration effect. Extensive experiments demonstrate the effectiveness of the proposed method. Our dataset is released at WIXRayNet.
****

MixFormer: Mixing Features Across Windows and Dimensions
Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5249-5259
While local-window self-attention performs notably in vision tasks, it suffers from limited receptive field and weak modeling capability issues. This is mainly because it performs self-attention within non-overlapped windows and shares weights on the channel dimension. We propose MixFormer to find a solution. First, we combine local-window self-attention with depth-wise convolution in a parallel design, modeling cross-window connections to enlarge the receptive fields. Second, we propose bi-directional interactions across branches to provide complementary clues in the channel and spatial dimensions. These two designs are integrated to achieve efficient feature mixing among windows and dimensions. Our MixFormer provides competitive results on image classification with EfficientNet and shows better results than RegNet and Swin Transformer. Performance in downstream tasks outperforms its alternatives by significant margins with less computational costs in 5 dense prediction tasks on MS COCO, ADE20k, and LVIS. Code is available at https://github.com/PaddlePaddle/PaddleClas.
****

Killing Two Birds With One Stone: Efficient and Robust Training of Face Recognition CNNs by Partial FC
Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, Tongliang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4042-4051
Learning discriminative deep feature embeddings by using million-scale in-the-wild datasets and margin-based softmax loss is the current state-of-the-art approach for face recognition. However, the memory and computing cost of the Fully Connected (FC) layer linearly scales up to the number of identities in the training set. Besides, the large-scale training data inevitably suffers from inter-class conflict and long-tailed distribution. In this paper, we propose a sparsely updating variant of the FC layer, named Partial FC (PFC). In each iteration, positive class centers and a random subset of negative class centers are selected to compute the margin-based softmax loss. All class centers are still maintained throughout the whole training process, but only a subset is selected and updated in each iteration. Therefore, the computing requirement, the probability of inter-class conflict, and the frequency of passive update on tail class centers, are dramatically reduced. Extensive experiments across different training data and backbones (e.g. CNN and ViT) confirm the effectiveness, robustness and efficiency of the proposed PFC.
****

NeRF-Editing: Geometry Editing of Neural Radiance Fields
Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, Lin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18353-18364
Implicit neural rendering, especially Neural Radiance Field (NeRF), has shown great potential in novel view synthesis of a scene. However, current NeRF-based methods cannot enable users to perform user-controlled shape deformation in the scene. While existing works have proposed some approaches to modify the radiance f

ield according to the user's constraints, the modification is limited to color e diting or object translation and rotation. In this paper, we propose a method th at allows users to perform controllable shape deformation on the implicit repres entation of the scene, and synthesizes the novel view images of the edited scene without re-training the network. Specifically, we establish a correspondence be tween the extracted explicit mesh representation and the implicit neural represe ntation of the target scene. Users can first utilize well-developed mesh-based d eformation methods to deform the mesh representation of the scene. Our method th en utilizes user edits from the mesh representation to bend the camera rays by i ntroducing a tetrahedra mesh as a proxy, obtaining the rendering results of the edited scene. Extensive experiments demonstrate that our framework can achieve i deal editing results not only on synthetic data, but also on real scenes capture d by users.

****

Optimal Correction Cost for Object Detection Evaluation

Mayu Otani, Riku Togashi, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Shin'ichi S atoh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2022, pp. 21107-21115

Mean Average Precision (mAP) is the primary evaluation measure for object detect ion. Although object detection has a broad range of applications, mAP evaluates detectors in terms of the performance of ranked instance retrieval. Such the ass umption for the evaluation task does not suit some downstream tasks. To alleviat e the gap between downstream tasks and the evaluation scenario, we propose Optim al Correction Cost (OC-cost), which assesses detection accuracy at image level. OC-cost computes the cost of correcting detections to ground truths as a measure of accuracy. The cost is obtained by solving an optimal transportation problem between the detections and the ground truths. Unlike mAP, OC-cost is designed to penalize false positive and false negative detections properly, and every image in a dataset is treated equally. Our experimental result validates that OC-cost has better agreement with human preference than a ranking-based measure, i.e., mAP for a single image. We also show that detectors' rankings by OC-cost are mor e consistent on different data splits than mAP. Our goal is not to replace mAP w ith OC-cost but provide an additional tool to evaluate detectors from another as pect. To help future researchers and developers choose a target measure, we prov ide a series of experiments to clarify how mAP and OC-cost differ.

****

Contextual Similarity Distillation for Asymmetric Image Retrieval

Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, Qi Tian; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9489-9 498

Asymmetric image retrieval, which typically uses small model for query side and large model for database server, is an effective solution for resource-constrain ed scenarios. However, existing approaches either fail to achieve feature cohere nce or make strong assumptions, e.g., requiring labeled datasets or classifiers from large model, etc., which limits their practical application. To this end, w e propose a flexible contextual similarity distillation framework to enhance the small query model and keep its output feature compatible with that of large gal lery model, which is crucial with asymmetric retrieval. In our approach, we lear n the small model with a new contextual similarity consistency constraint withou t any data label. During the small model learning, it preserves the contextual s imilarity among each training image and its neighbors with the features extracte d by the large model. Note that this simple constraint is consistent with simult aneous first-order feature vector preserving and second-order ranking list prese rving. Extensive experiments show that the proposed method outperforms the state -of-the-art methods on the Revisited Oxford and Paris datasets.

****

FineDiving: A Fine-Grained Dataset for Procedure-Aware Action Quality Assessment

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, Jiwen Lu; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2949-2958

Most existing action quality assessment methods rely on the deep features of an entire video to predict the score, which is less reliable due to the non-transparent inference process and poor interpretability. We argue that understanding both high-level semantics and internal temporal structures of actions in competitive sports videos is the key to making predictions accurate and interpretable. Towards this goal, we construct a new fine-grained dataset, called FineDiving, developed on diverse diving events with detailed annotations on action procedures. We also propose a procedure-aware approach for action quality assessment, learned by a new Temporal Segmentation Attention module. Specifically, we propose to parse pairwise query and exemplar action instances into consecutive steps with diverse semantic and temporal correspondences. The procedure-aware cross-attention is proposed to learn embeddings between query and exemplar steps to discover their semantic, spatial, and temporal correspondences, and further serve for fine-grained contrastive regression to derive a reliable scoring mechanism. Extensive experiments demonstrate that our approach achieves substantial improvements over the state-of-the-art methods with better interpretability. The dataset and code are available at https://github.com/xujinglin/FineDiving.

****

Artistic Style Discovery With Independent Components

Xin Xie, Yi Li, Huaibo Huang, Haiyan Fu, Wanwan Wang, Yanqing Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19870-19879

Style transfer has been well studied in recent years with excellent performance processed. While existing methods usually choose CNNs as the powerful tool to accomplish superb stylization, less attention was paid to the latent style space. Rare exploration of underlying dimensions results in the poor style controllability and the limited practical application. In this work, we rethink the internal meaning of style features, further proposing a novel unsupervised algorithm for style discovery and achieving personalized manipulation. In particular, we take a closer look into the mechanism of style transfer and obtain different artistic style components from the latent space consisting of different style features. Then fresh styles can be generated by linear combination according to various style components. Experimental results have shown that our approach is superb in 1) restylizing the original output with the diverse artistic styles discovered from the latent space while keeping the content unchanged, and 2) being generic and compatible for various style transfer methods.

****

HEAT: Holistic Edge Attention Transformer for Structured Reconstruction

Jiacheng Chen, Yiming Qian, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3866-3875

This paper presents a novel attention-based neural network for structured reconstruction, which takes a 2D raster image as an input and reconstructs a planar graph depicting an underlying geometric structure. The approach detects corners and classifies edge candidates between corners in an end-to-end manner. Our contribution is a holistic edge classification architecture, which 1) initializes the feature of an edge candidate by a trigonometric positional encoding of its end-points; 2) fuses image feature to each edge candidate by deformable attention; 3) employs two weight-sharing Transformer decoders to learn holistic structural patterns over the graph edge candidates; and 4) is trained with a masked learning strategy. The corner detector is a variant of the edge classification architecture, adapted to operate on pixels as corner candidates. We conduct experiments on two structured reconstruction tasks: outdoor building architecture and indoor floorplan planar graph reconstruction. Extensive qualitative and quantitative evaluations demonstrate the superiority of our approach over the state of the art. Code and pre-trained models are available at https://heat-structured-reconstruction.github.io

****

HyperStyle: StyleGAN Inversion With HyperNetworks for Real Image Editing

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, Amit Bermano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp.

18511-18521
The inversion of real images into StyleGAN's latent space is a well-studied problem. Nevertheless, applying existing approaches to real-world scenarios remains an open challenge, due to an inherent trade-off between reconstruction and editability: latent space regions which can accurately represent real images typically suffer from degraded semantic control. Recent work proposes to mitigate this trade-off by fine-tuning the generator to add the target image to well-behaved, editable regions of the latent space. While promising, this fine-tuning scheme is impractical for prevalent use as it requires a lengthy training phase for each new image. In this work, we introduce this approach into the realm of encoder-based inversion. We propose HyperStyle, a hypernetwork that learns to modulate StyleGAN's weights to faithfully express a given image in editable regions of the latent space. A naive modulation approach would require training a hypernetwork with over three billion parameters. Through careful network design, we reduce this to be in line with existing encoders. HyperStyle yields reconstructions comparable to those of optimization techniques with the near real-time inference capabilities of encoders. Lastly, we demonstrate HyperStyle's effectiveness on several applications beyond the inversion task, including the editing of out-of-domain images which were never seen during training.
****
DASO: Distribution-Aware Semantics-Oriented Pseudo-Label for Imbalanced Semi-Supervised Learning
Youngtaek Oh, Dong-Jin Kim, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9786-9796
The capability of the traditional semi-supervised learning (SSL) methods is far from real-world application due to severely biased pseudo-labels caused by (1) class imbalance and (2) class distribution mismatch between labeled and unlabeled data. This paper addresses such a relatively under-explored problem. First, we propose a general pseudo-labeling framework that class-adaptively blends the semantic pseudo-label from a similarity-based classifier to the linear one from the linear classifier, after making the observation that both types of pseudo-labels have complementary properties in terms of bias. We further introduce a novel semantic alignment loss to establish balanced feature representation to reduce the biased predictions from the classifier. We term the whole framework as Distribution-Aware Semantics-Oriented (DASO) Pseudo-label. We conduct extensive experiments in a wide range of imbalanced benchmarks: CIFAR10/100-LT, STL10-LT, and large-scale long-tailed Semi-Aves with open-set class, and demonstrate that, the proposed DASO framework reliably improves SSL learners with unlabeled data especially when both (1) class imbalance and (2) distribution mismatch dominate.
****
Mobile-Former: Bridging MobileNet and Transformer
Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5270-5279
We present Mobile-Former, a parallel design of MobileNet and transformer with a two-way bridge in between. This structure leverages the advantages of MobileNet at local processing and transformer at global interaction. And the bridge enables bidirectional fusion of local and global features. Different from recent works on vision transformer, the transformer in Mobile-Former contains very few tokens (e.g. 6 or fewer tokens) that are randomly initialized to learn global priors, resulting in low computational cost. Combining with the proposed light-weight cross attention to model the bridge, Mobile-Former is not only computationally efficient, but also has more representation power. It outperforms MobileNetV3 at low FLOP regime from 25M to 500M FLOPs on ImageNet classification. For instance, Mobile-Former achieves 77.9% top-1 accuracy at 294M FLOPs, gaining 1.3% over MobileNetV3 but saving 17% of computations. When transferring to object detection, Mobile-Former outperforms MobileNetV3 by 8.6 AP in RetinaNet framework. Furthermore, we build an efficient end-to-end detector by replacing backbone, encoder and decoder in DETR with Mobile-Former, which outperforms DETR by 1.3 AP but saves 52% of computational cost and 36% of parameters. Code will be released at https

://github.com/aaboys/mobileformer.
****

Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation
Andra Petrovai, Sergiu Nedevschi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1578-1588
We present a novel self-distillation based self-supervised monocular depth estimation (SD-SSMDE) learning framework. In the first step, our network is trained in a self-supervised regime on high-resolution images with the photometric loss. The network is further used to generate pseudo depth labels for all the images in the training set. To improve the performance of our estimates, in the second step, we re-train the network with the scale invariant logarithmic loss supervised by pseudo labels. We resolve scale ambiguity and inter-frame scale consistency by introducing an automatically computed scale in our depth labels. To filter out noisy depth values, we devise a filtering scheme based on the 3D consistency between consecutive views. Extensive experiments demonstrate that each proposed component and the self-supervised learning framework improve the quality of the depth estimation over the baseline and achieve state-of-the-art results on the KITTI and Cityscapes datasets.
****

DESTR: Object Detection With Split Transformer
Liqiang He, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9377-9386
Self- and cross-attention in Transformers provide for high model capacity, making them viable models for object detection. However, Transformers still lag in performance behind CNN-based detectors. This is, we believe, because: (a) Cross-attention is used for both classification and bounding-box regression tasks; (b) Transformer's decoder poorly initializes content queries; and (c) Self-attention poorly accounts for certain prior knowledge which could help improve inductive bias. These limitations are addressed with the corresponding three contributions. First, we propose a new Detection Split Transformer (DESTR) that separates estimation of cross-attention into two independent branches -- one tailored for classification and the other for box regression. Second, we use a mini-detector to initialize the content queries in the decoder with classification and regression embeddings of the respective heads in the mini-detector. Third, we augment self-attention in the decoder to additionally account for pairs of adjacent object queries. Our experiments on the MS-COCO dataset show that DESTR outperforms DETR and its successors.
****

LTP: Lane-Based Trajectory Prediction for Autonomous Driving
Jingke Wang, Tengju Ye, Ziqing Gu, Junbo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17134-17142
The reasonable trajectory prediction of surrounding traffic participants is crucial for autonomous driving. Especially, how to predict multiple plausible trajectories is still a challenging problem because of the multiple possibilities of the future. Proposal-based prediction methods address the multi-modality issues with a two-stage approach, commonly using intention classification followed by motion regression. This paper proposes a two-stage proposal-based motion forecasting method that exploits the sliced lane segments as fine-grained, shareable, and interpretable proposals. We use Graph neural network and Transformer to encode the shape and interaction information among the map sub-graphs and the agents sub-graphs. In addition, we propose a variance-based non-maximum suppression strategy to select representative trajectories that ensure the diversity of the final output. Experiments on the Argoverse dataset show that the proposed method outperforms state-of-the-art methods, and the lane segments-based proposals as well as the variance-based non-maximum suppression strategy both contribute to the performance improvement. Moreover, we demonstrate that the proposed method can achieve reliable performance with a lower collision rate and fewer off-road scenarios in the closed-loop simulation.
****

CycleMix: A Holistic Strategy for Medical Image Segmentation From Scribble Supervision
Ke Zhang, Xiahai Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11656-11665
Curating a large set of fully annotated training data can be costly, especially for the tasks of medical image segmentation. Scribble, a weaker form of annotation, is more obtainable in practice, but training segmentation models from limited supervision of scribbles is still challenging. To address the difficulties, we propose a new framework for scribble learning-based medical image segmentation, which is composed of mix augmentation and cycle consistency and thus is referred to as CycleMix. For augmentation of supervision, CycleMix adopts the mixup strategy with a dedicated design of random occlusion, to perform increments and decrements of scribbles. For regularization of supervision, CycleMix intensifies the training objective with consistency losses to penalize inconsistent segmentation, which results in significant improvement of segmentation performance. Results on two open datasets, i.e., ACDC and MSCMRseg, showed that the proposed method achieved exhilarating performance, demonstrating comparable or even better accuracy than the fully-supervised methods. The code and expert-made scribble annotations for MSCMRseg are publicly available at https://github.com/BWGZK/CycleMIx.
****
VideoINR: Learning Video Implicit Neural Representation for Continuous Space-Time Super-Resolution
Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2047-2057
Videos typically record the streaming and continuous visual data as discrete consecutive frames. Since the storage cost is expensive for videos of high fidelity, most of them are stored in a relatively low resolution and frame rate. Recent works of Space-Time Video Super-Resolution (STVSR) are developed to incorporate temporal interpolation and spatial super-resolution in a unified framework. However, most of them only support a fixed up-sampling scale, which limits their flexibility and applications. In this work, instead of following the discrete representations, we propose Video Implicit Neural Representation (VideoINR), and we show its applications for STVSR. The learned implicit neural representation can be decoded to videos of arbitrary spatial resolution and frame rate. We show that VideoINR achieves competitive performances with state-of-the-art STVSR methods on common up-sampling scales and significantly outperforms prior works on continuous and out-of-training-distribution scales. Our project page is at http://zeyuan-chen.com/VideoINR/ and code is available at https://github.com/Picsart-AI-Research/VideoINR-Continuous-Space-Time-Super-Resolution.
****
Towards End-to-End Unified Scene Text Detection and Layout Analysis
Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, Michalis Raptis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1049-1059
Scene text detection and document layout analysis have long been treated as two separate tasks in different image domains. In this paper, we bring them together and introduce the task of unified scene text detection and layout analysis. The first hierarchical scene text dataset is introduced to enable this novel research task. We also propose a novel method that is able to simultaneously detect scene text and form text clusters in a unified way. Comprehensive experiments show that our unified model achieves better performance than multiple well-designed baseline methods. Additionally, this model achieves state-of-the-art results on multiple scene text detection datasets without the need of complex post-processing. Dataset and code: https://github.com/google-research-datasets/hiertext.
****
Image Based Reconstruction of Liquids From 2D Surface Detections
Florian Richter, Ryan K. Orosco, Michael C. Yip; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13811-13820
In this work, we present a solution to the challenging problem of reconstructing

liquids from image data. The challenges in reconstructing liquids, which is not faced in previous reconstruction works on rigid and deforming surfaces, lies in the inability to use depth sensing and color features due the variable index of refraction, opacity, and environmental reflections. Therefore, we limit ourselves to only surface detections (i.e. binary mask) of liquids as observations and do not assume any prior knowledge on the liquids properties. A novel optimization problem is posed which reconstructs the liquid as particles by minimizing the error between a rendered surface from the particles and the surface detections while satisfying liquid constraints. Our solvers to this optimization problem are presented and no training data is required to apply them. We also propose a dynamic prediction to seed the reconstruction optimization from the previous time-step. We test our proposed methods in simulation and on two new liquid datasets which we open source so the broader research community can continue developing in this under explored area.

****

Contextual Outpainting With Object-Level Contrastive Learning

Jiacheng Li, Chang Chen, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11451-11460

We study the problem of contextual outpainting, which aims to hallucinate the missing background contents based on the remaining foreground contents. Existing image outpainting methods focus on completing object shapes or extending existing scenery textures, neglecting the semantically meaningful relationship between the missing and remaining contents. To explore the semantic cues provided by the remaining foreground contents, we propose a novel ConTextual Outpainting GAN (CTO-GAN), leveraging the semantic layout as a bridge to synthesize coherent and diverse background contents. To model the contextual correlation between foreground and background contents, we incorporate an object-level contrastive loss to regularize the learning of cross-modal representations of foreground contents and the corresponding background semantic layout, facilitating accurate semantic reasoning. Furthermore, we improve the realism of the generated background contents via detecting generated context in adversarial training. Extensive experiments demonstrate that the proposed method achieves superior performance compared with existing solutions on the challenging COCO-stuff dataset.

****

AP-BSN: Self-Supervised Denoising for Real-World Images via Asymmetric PD and Blind-Spot Network

Wooseok Lee, Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17725-17734

Blind-spot network (BSN) and its variants have made significant advances in self-supervised denoising. Nevertheless, they are still bound to synthetic noisy inputs due to less practical assumptions like pixel-wise independent noise. Hence, it is challenging to deal with spatially correlated real-world noise using self-supervised BSN. Recently, pixel-shuffle downsampling (PD) has been proposed to remove the spatial correlation of real-world noise. However, it is not trivial to integrate PD and BSN directly, which prevents the fully self-supervised denoising model on real-world images. We propose an Asymmetric PD (AP) to address this issue, which introduces different PD stride factors for training and inference. We systematically demonstrate that the proposed AP can resolve inherent trade-offs caused by specific PD stride factors and make BSN applicable to practical scenarios. To this end, we develop AP-BSN, a state-of-the-art self-supervised denoising method for real-world sRGB images. We further propose random-replacing refinement, which significantly improves the performance of our AP-BSN without any additional parameters. Extensive studies demonstrate that our method outperforms the other self-supervised and even unpaired denoising methods by a large margin, without using any additional knowledge, e.g., noise level, regarding the underlying unknown noise.

****

AutoSDF: Shape Priors for 3D Completion, Reconstruction and Generation

Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, Shubham Tulsiani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,

pp. 306-315
Powerful priors allow us to perform inference with insufficient information. In this paper, we propose an autoregressive prior for 3D shapes to solve multimodal 3D tasks such as shape completion, reconstruction, and generation. We model the distribution over 3D shapes as a non-sequential autoregressive distribution over a discretized, low-dimensional, symbolic grid-like latent representation of 3D shapes. This enables us to represent distributions over 3D shapes conditioned on information from an arbitrary set of spatially anchored query locations and thus perform shape completion in such arbitrary settings (e.g. generating a complete chair given only a view of the back leg). We also show that the learned autoregressive prior can be leveraged for conditional tasks such as single-view reconstruction and language-based generation. This is achieved by learning task-specific 'naive' conditionals which can be approximated by light-weight models trained on minimal paired data. We validate the effectiveness of the proposed method using both quantitative and qualitative evaluation and show that the proposed method outperforms the specialized state-of-the-art methods trained for individual tasks. The project page with code and video visualizations can be found at https://yccyenchicheng.github.io/AutoSDF/.

****

ISNAS-DIP: Image-Specific Neural Architecture Search for Deep Image Prior
Metin Ersin Arican, Ozgur Kara, Gustav Bredell, Ender Konukoglu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1960-1968
Recent works show that convolutional neural network (CNN) architectures have a spectral bias towards lower frequencies, which has been leveraged for various image restoration tasks in the Deep Image Prior (DIP) framework. The benefit of the inductive bias the network imposes in the DIP framework depends on the architecture. Therefore, researchers have studied how to automate the search to determine the best-performing model. However, common neural architecture search (NAS) techniques are resource and time-intensive. Moreover, best-performing models are determined for a whole dataset of images instead of for each image independently, which would be prohibitively expensive. In this work, we first show that optimal neural architectures in the DIP framework are image-dependent. Leveraging this insight, we then propose an image-specific NAS strategy for the DIP framework that requires substantially less training than typical NAS approaches, effectively enabling image-specific NAS. We justify the proposed strategy's effectiveness by (1) demonstrating its performance on a NAS Dataset for DIP that includes 522 models from a particular search space (2) conducting extensive experiments on image denoising, inpainting, and super-resolution tasks. Our experiments show that image-specific metrics can reduce the search space to a small cohort of models, of which the best model outperforms current NAS approaches for image restoration. Codes and datasets are available at https://github.com/ozgurkara99/ISNAS-DIP.

****

Depth-Guided Sparse Structure-From-Motion for Movies and TV Shows
Sheng Liu, Xiaohan Nie, Raffay Hamid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15980-15989
Existing approaches for Structure from Motion (SfM) produce impressive 3D reconstruction results especially when using imagery captured with large parallax. However, to create engaging video-content in movies and TV shows, the amount by which a camera can be moved while filming a particular shot is often limited. The resulting small-motion parallax between video frames makes standard geometry-based SfM approaches not as effective for movies and TV shows. To address this challenge, we propose a simple yet effective approach that uses single-frame depth-prior obtained from a pretrained network to significantly improve geometry-based SfM for our small-parallax setting. To this end, we first use the depth-estimates of the detected keypoints to reconstruct the point cloud and camera-pose for initial two-view reconstruction. We then perform depth-regularized optimization to register new images and triangulate the new points during incremental reconstruction. To comprehensively evaluate our approach, we introduce a new dataset (StudioSfM) consisting of 130 shots with 21K frames from 15 studio-produced videos t

hat are manually annotated by a professional CG studio. We demonstrate that our approach: (a) significantly improves the quality of 3D reconstruction for our small-parallax setting, (b) does not cause any degradation for data with large-parallax, and (c) maintains the generalizability and scalability of geometry-based sparse SfM. Our dataset can be obtained at github.com/amazon-research/small-baseline-camera-tracking.
****

End-to-End Referring Video Object Segmentation With Multimodal Transformers
Adam Botach, Evgenii Zheltonozhskii, Chaim Baskin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4985-4995
The referring video object segmentation task (RVOS) involves segmentation of a text-referred object instance in the frames of a given video. Due to the complex nature of this multimodal task, which combines text reasoning, video understanding, instance segmentation and tracking, existing approaches typically rely on sophisticated pipelines in order to tackle it. In this paper, we propose a simple Transformer-based approach to RVOS. Our framework, termed Multimodal Tracking Transformer (MTTR), models the RVOS task as a sequence prediction problem. Following recent advancements in computer vision and natural language processing, MTTR is based on the realization that video and text can be processed together effectively and elegantly by a single multimodal Transformer model. MTTR is end-to-end trainable, free of text-related inductive bias components and requires no additional mask-refinement post-processing steps. As such, it simplifies the RVOS pipeline considerably compared to existing methods. Evaluation on standard benchmarks reveals that MTTR significantly outperforms previous art across multiple metrics. In particular, MTTR shows impressive +5.7 and +5.0 mAP gains on the A2D-Sentences and JHMDB-Sentences datasets respectively, while processing 76 frames per second. In addition, we report strong results on the public validation set of Refer-YouTube-VOS, a more challenging RVOS dataset that has yet to receive the attention of researchers. The code to reproduce our experiments is available at https://github.com/mttr2021/MTTR
****

Unpaired Cartoon Image Synthesis via Gated Cycle Mapping
Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, Xuansong Xie, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3501-3510
In this paper, we present a general-purpose solution to cartoon image synthesis with unpaired training data. In contrast to previous works learning pre-defined cartoon styles for specified usage scenarios (portrait or scene), we aim to train a common cartoon translator which can not only simultaneously render exaggerated anime faces and realistic cartoon scenes, but also provide flexible user controls for desired cartoon styles. It is challenging due to the complexity of the task and the absence of paired data. The core idea of the proposed method is to introduce gated cycle mapping, that utilizes a novel gated mapping unit to produce the category-specific style code and embeds this code into cycle networks to control the translation process. For the concept of category, we classify images into different categories (e.g., 4 types: photo/cartoon portrait/scene) and learn finer-grained category translations rather than overall mappings between two domains (e.g., photo and cartoon). Furthermore, the proposed method can be easily extended to cartoon video generation with an auxiliary dataset and a new adaptive style loss. Experimental results demonstrate the superiority of the proposed method over the state of the art and validate its effectiveness in the brand-new task of general cartoon image synthesis.
****

IterMVS: Iterative Probability Estimation for Efficient Multi-View Stereo
Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8606-8615
We present IterMVS, a new data-driven method for high-resolution multi-view stereo. We propose a novel GRU-based estimator that encodes pixel-wise probability distributions of depth in its hidden state. Ingesting multi-scale matching inform

ation, our model refines these distributions over multiple iterations and infers depth and confidence. To extract the depth maps, we combine traditional classification and regression in a novel manner. We verify the efficiency and effectiveness of our method on DTU, Tanks&Temples and ETH3D. While being the most efficient method in both memory and run-time, our model achieves competitive performance on DTU and better generalization ability on Tanks&Temples as well as ETH3D than most state-of-the-art methods. Code is available at https://github.com/FangjinhuaWang/IterMVS.
****

Not All Points Are Equal: Learning Highly Efficient Point-Based Detectors for 3D LiDAR Point Clouds
Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18953-18962
We study the problem of efficient object detection of 3D LiDAR point clouds. To reduce the memory and computational cost, existing point-based pipelines usually adopt task-agnostic random sampling or farthest point sampling to progressively downsample input point clouds, despite the fact that not all points are equally important to the task of object detection. In particular, the foreground points are inherently more important than background points for object detectors. Motivated by this, we propose a highly-efficient single-stage point-based 3D detector in this paper, termed IA-SSD. The key of our approach is to exploit two learnable, task-oriented, instance-aware downsampling strategies to hierarchically select the foreground points belonging to objects of interest. Additionally, we also introduce a contextual centroid perception module to further estimate precise instance centers. Finally, we build our \nickname following the encoder-only architecture for efficiency. Extensive experiments conducted on several large-scale detection benchmarks demonstrate the competitive performance of our IA-SSD. Thanks to the low memory footprint and a high degree of parallelism, it achieves a superior speed of 80+ frames-per-second on the KITTI dataset with a single RTX 2080Ti GPU. The code is available at https://github.com/yifanzhang713/IA-SSD.
****

FedCorr: Multi-Stage Federated Learning for Label Noise Correction
Jingyi Xu, Zihan Chen, Tony Q.S. Quek, Kai Fong Ernest Chong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10184-10193
Federated learning (FL) is a privacy-preserving distributed learning paradigm that enables clients to jointly train a global model. In real-world FL implementations, client data could have label noise, and different clients could have vastly different label noise levels. Although there exist methods in centralized learning for tackling label noise, such methods do not perform well on heterogeneous label noise in FL settings, due to the typically smaller sizes of client datasets and data privacy requirements in FL. In this paper, we propose FedCorr, a general multi-stage framework to tackle heterogeneous label noise in FL, without making any assumptions on the noise models of local clients, while still maintaining client data privacy. In particular, (1) FedCorr dynamically identifies noisy clients by exploiting the dimensionalities of the model prediction subspaces independently measured on all clients, and then identifies incorrect labels on noisy clients based on per-sample losses. To deal with data heterogeneity and to increase training stability, we propose an adaptive local proximal regularization term that is based on estimated local noise levels. (2) We further finetune the global model on identified clean clients and correct the noisy labels for the remaining noisy clients after finetuning. (3) Finally, we apply the usual training on all clients to make full use of all local data. Experiments conducted on CIFAR-10/100 with federated synthetic label noise, and on a real-world noisy dataset, Clothing1M, demonstrate that FedCorr is robust to label noise and substantially outperforms the state-of-the-art methods at multiple noise levels.
****

Detecting Camouflaged Object in Frequency Domain
Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, Shouhong Ding; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4504-4513
Camouflaged object detection (COD) aims to identify objects that are perfectly embedded in their environment, which has various downstream applications in fields such as medicine, art, and agriculture. However, it is an extremely challenging task to spot camouflaged objects with the perception ability of human eyes. Hence, we claim that the goal of COD task is not just to mimic the human visual ability in a single RGB domain, but to go beyond the human biological vision. We then introduce the frequency domain as an additional clue to better detect camouflaged objects from backgrounds. To well involve the frequency clues into the CNN models, we present a powerful network with two special components. We first design a novel frequency enhancement module (FEM) to dig clues of camouflaged objects in the frequency domain. It contains the offline discrete cosine transform followed by the learnable enhancement. Then we use a feature alignment to fuse the features from RGB domain and frequency domain. Moreover, to further make full use of the frequency information, we propose the high-order relation module (HOR) to handle the rich fusion feature. Comprehensive experiments on three widely-used COD datasets show the proposed method significantly outperforms other state-of-the-art methods by a large margin. The code and results are released in https://github.com/luckybird1994/FDCOD.
****

RigNeRF: Fully Controllable Neural 3D Portraits
ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, Zhixin Shu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20364-20373
Volumetric neural rendering methods, such as neural ra-diance fields (NeRFs), have enabled photo-realistic novel view synthesis. However, in their standard form, NeRFs do not support the editing of objects, such as a human head,within a scene. In this work, we propose RigNeRF, a system that goes beyond just novel view synthesis and enables full control of head pose and facial expressions learned from a single portrait video. We model changes in head pose and facial expressions using a deformation field that is guided by a 3D morphable face model (3DMM). The 3DMM effectively acts as a prior for RigNeRF that learns to predict only residuals to the 3DMM deformations and allows us to render novel (rigid) poses and (non-rigid) expressions that were not present in the input sequence. Using only a smartphone-captured short video of a subject for training,we demonstrate the effectiveness of our method on free view synthesis of a portrait scene with explicit head pose and expression controls.
****

CLIP-Forge: Towards Zero-Shot Text-To-Shape Generation
Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, Kamal Rahimi Malekshan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18603-18613
Generating shapes using natural language can enable new ways of imagining and creating the things around us. While significant recent progress has been made in text-to-image generation, text-to-shape generation remains a challenging problem due to the unavailability of paired text and shape data at a large scale. We present a simple yet effective method for zero-shot text-to-shape generation that circumvents such data scarcity. Our proposed method, named CLIP-Forge, is based on a two-stage training process, which only depends on an unlabelled shape dataset and a pre-trained image-text network such as CLIP. Our method has the benefits of avoiding expensive inference time optimization, as well as the ability to generate multiple shapes for a given text. We not only demonstrate promising zero-shot generalization of the CLIP-Forge model qualitatively and quantitatively, but also provide extensive comparative evaluations to better understand its behavior.
****

Style-Based Global Appearance Flow for Virtual Try-On
Sen He, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3470-3479

Image-based virtual try-on aims to fit an in-shop garment into a clothed person image. To achieve this, a key step is garment warping which spatially aligns the target garment with the corresponding body parts in the person image. Prior methods typically adopt a local appearance flow estimation model. They are thus intrinsically susceptible to difficult body poses/occlusions and large mis-alignments between person and garment images. To overcome this limitation, a novel global appearance flow estimation model is proposed in this work. For the first time, a StyleGAN based architecture is adopted for appearance flow estimation. This enables us to take advantage of a global style vector to encode a whole-image context to cope with the aforementioned challenges. To guide the StyleGAN flow generator to pay more attention to local garment deformation, a flow refinement module is introduced to add local context. Experiment results on a popular virtual try-on benchmark show that our method achieves new state-of-the-art performance. It is particularly effective in a 'in-the-wild' application scenario where the reference image is full-body resulting in a large mis-alignment with the garment image.

****

Source-Free Object Detection by Learning To Overlook Domain Style

Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, Lin Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8014-8023

Source-free object detection (SFOD) needs to adapt a detector pre-trained on a labeled source domain to a target domain, with only unlabeled training data from the target domain. Existing SFOD methods typically adopt the pseudo labeling paradigm with model adaption alternating between predicting pseudo labels and fine-tuning the model. This approach suffers from both unsatisfactory accuracy of pseudo labels due to the presence of domain shift and limited use of target domain training data. In this work, we present a novel Learning to Overlook Domain Style (LODS) method with such limitations solved in a principled manner. Our idea is to reduce the domain shift effect by enforcing the model to overlook the target domain style, such that model adaptation is simplified and becomes easier to carry on. To that end, we enhance the style of each target domain image and leverage the style degree difference between the original image and the enhanced image as a self-supervised signal for model adaptation. By treating the enhanced image as an auxiliary view, we exploit a student-teacher architecture for learning to overlook the style degree difference against the original image, also characterized with a novel style enhancement algorithm and graph alignment constraint. Extensive experiments demonstrate that our LODS yields new state-of-the-art performance on four benchmarks.

****

Active Learning for Open-Set Annotation

Kun-Peng Ning, Xun Zhao, Yu Li, Sheng-Jun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 41-49

Existing active learning studies typically work in the closed-set setting by assuming that all data examples to be labeled are drawn from known classes. However, in real annotation tasks, the unlabeled data usually contains a large amount of examples from unknown classes, resulting in the failure of most active learning methods. To tackle this open-set annotation (OSA) problem, we propose a new active learning framework called LfOSA, which boosts the classification performance with an effective sampling strategy to precisely detect examples from known classes for annotation. The LfOSA framework introduces an auxiliary network to model the per-example max activation value (MAV) distribution with a Gaussian Mixture Model, which can dynamically select the examples with highest probability from known classes in the unlabeled set. Moreover, by reducing the temperature T of the loss function, the detection model will be further optimized by exploiting both known and unknown supervision. The experimental results show that the proposed method can significantly improve the selection quality of known classes, and achieve higher classification accuracy with lower annotation cost than state-of-the-art active learning methods. To the best of our knowledge, this is the first work of active learning for open-set annotation.

****
SceneSqueezer: Learning To Compress Scene for Camera Relocalization
Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8259-8268
Standard visual localization methods build a priori 3D model of a scene which is used to establish correspondences against the 2D keypoints in a query image. Storing these pre-built 3D scene models can be prohibitively expensive for large-scale environments, especially on mobile devices with limited storage and communication bandwidth. We design a novel framework that compresses a scene while still maintaining localization accuracy. The scene is compressed in three stages: first, the database frames are clustered using pairwise co-visibility information. Then, a learned point selection module prunes the points in each cluster taking into account the final pose estimation accuracy. In the final stage, the features of the selected points are further compressed using learned quantization. Query image registration is done using only the compressed scene points. To the best of our knowledge, we are the first to propose learned scene compression for visual localization. We also demonstrate the effectiveness and efficiency of our method on various outdoor datasets where it can perform accurate localization with low memory consumption.
****
SelfRecon: Self Reconstruction Your Digital Avatar From Monocular Video
Boyi Jiang, Yang Hong, Hujun Bao, Juyong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5605-5615
We propose SelfRecon, a clothed human body reconstruction method that combines implicit and explicit representations to recover space-time coherent geometries from a monocular self-rotating human video. Explicit methods require a predefined template mesh for a given sequence, while the template is hard to acquire for a specific subject. Meanwhile, the fixed topology limits the reconstruction accuracy and clothing types. Implicit representation supports arbitrary topology and can represent high-fidelity geometry shapes due to its continuous nature. However, it is difficult to integrate multi-frame information to produce a consistent registration sequence for downstream applications. We propose to combine the advantages of both representations. We utilize differential mask loss of the explicit mesh to obtain the coherent overall shape, while the details on the implicit surface are refined with the differentiable neural rendering. Meanwhile, the explicit mesh is updated periodically to adjust its topology changes, and a consistency loss is designed to match both representations. Compared with existing methods, SelfRecon can produce high-fidelity surfaces for arbitrary clothed humans with self-supervised optimization. Extensive experimental results demonstrate its effectiveness on real captured monocular videos. The source code is available at https://github.com/jby1993/SelfReconCode.
****
Instance-Dependent Label-Noise Learning With Manifold-Regularized Transition Matrix Estimation
De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, Masashi Sugiyama; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16630-16639
In label-noise learning, estimating the transition matrix has attracted more and more attention as the matrix plays an important role in building statistically consistent classifiers. However, it is very challenging to estimate the transition matrix T(x), where T(x) denotes the instance, because it is unidentifiable under the instance-dependent noise (IDN). To address this problem, we have noticed that, there are psychological and physiological evidences showing that we humans are more likely to annotate instances of similar appearances to the same classes, and thus poor-quality or ambiguous instances of similar appearances are easier to be mislabeled to the correlated or same noisy classes. Therefore, we propose assumption on the geometry of T(x) that "the closer two instances are, the more similar their corresponding transition matrices should be". More specifically, we formulate above assumption into the manifold embedding, to effectively redu

ce the degree of freedom of T(x) and make it stably estimable in practice. This proposed manifold-regularized technique works by directly reducing the estimation error without hurting the approximation error about the estimation problem of T(x) Experimental evaluations on four synthetic and two real-world datasets demonstrate our method is superior to state-of-the-art approaches for label-noise learning under the challenging IDN.

****

Rethinking the Augmentation Module in Contrastive Learning: Learning Hierarchical Augmentation Invariance With Expanded Views

Junbo Zhang, Kaisheng Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16650-16659

A data augmentation module is utilized in contrastive learning to transform the given data example into two views, which is considered essential and irreplaceable. However, the pre-determined composition of multiple data augmentations brings two drawbacks. First, the artificial choice of augmentation types brings specific representational invariances to the model, which have different degrees of positive and negative effects on different downstream tasks. Treating each type of augmentation equally during training makes the model learn non-optimal representations for various downstream tasks and limits the flexibility to choose augmentation types beforehand. Second, the strong data augmentations used in classic contrastive learning methods may bring too much invariance in some cases, and fine-grained information that is essential to some downstream tasks may be lost. This paper proposes a general method to alleviate these two problems by considering "where" and "what" to contrast in a general contrastive learning framework. We first propose to learn different augmentation invariances at different depths of the model according to the importance of each data augmentation instead of learning representational invariances evenly in the backbone. We then propose to expand the contrast content with augmentation embeddings to reduce the misleading effects of strong data augmentations. Experiments based on several baseline methods demonstrate that we learn better representations for various benchmarks on classification, detection, and segmentation downstream tasks.

****

Self-Supervised Models Are Continual Learners

Enrico Fini, Victor G. Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, Julien Mairal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9621-9630

Self-supervised models have been shown to produce comparable or better visual representations than their supervised counterparts when trained offline on unlabeled data at scale. However, their efficacy is catastrophically reduced in a Continual Learning (CL) scenario where data is presented to the model sequentially. In this paper, we show that self-supervised loss functions can be seamlessly converted into distillation mechanisms for CL by adding a predictor network that maps the current state of the representations to their past state. This enables us to devise a framework for Continual self-supervised visual representation Learning that (i) significantly improves the quality of the learned representations, (ii) is compatible with several state-of-the-art self-supervised objectives, and (iii) needs little to no hyperparameter tuning. We demonstrate the effectiveness of our approach empirically by training six popular self-supervised models in various CL settings. Code: github.com/DonkeyShot21/cassle

****

Dreaming To Prune Image Deraining Networks

Weiqi Zou, Yang Wang, Xueyang Fu, Yang Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6023-6032

Convolutional image deraining networks have achieved great success while suffering from tremendous computational and memory costs. Most model compression methods require original data for iterative fine-tuning, which is limited in real-world applications due to storage, privacy, and transmission constraints. We note that it is overstretched to fine-tune the compressed model using self-collected data, as it exhibits poor generalization over images with different degradation characteristics. To address this problem, we propose a novel data-free compression

framework for deraining networks. It is based on our observation that deep degradation representations can be clustered by degradation characteristics (types of rain) while independent of image content. Therefore, in our framework, we "dream" diverse in-distribution degraded images using a deep inversion paradigm, thus leveraging them to distill the pruned model. Specifically, we preserve the performance of the pruned model in a dual-branch way. In one branch, we invert the pre-trained model (teacher) to reconstruct the degraded inputs that resemble the original distribution and employ the orthogonal regularization for deep features to yield degradation diversity. In the other branch, the pruned model (student) is distilled to fit the teacher's original statistical modeling on these dreamed inputs. Further, an adaptive pruning scheme is proposed to determine the hierarchical sparsity, which alleviates the regression drift of the initial pruned model. Experiments on various deraining datasets demonstrate that our method can reduce about 40% FLOPs of the state-of-the-art models while maintaining comparable performance without original data.

****

Equivariant Point Cloud Analysis via Learning Orientations for Message Passing
Shitong Luo, Jiahan Li, Jiaqi Guan, Yufeng Su, Chaoran Cheng, Jian Peng, Jianzhu Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18932-18941

Equivariance has been a long-standing concern in various fields ranging from computer vision to physical modeling. Most previous methods struggle with generality, simplicity, and expressiveness --- some are designed ad hoc for specific data types, some are too complex to be accessible, and some sacrifice flexible transformations. In this work, we propose a novel and simple framework to achieve equivariance for point cloud analysis based on the message passing (graph neural network) scheme. We find the equivariant property could be obtained by introducing an orientation for each point to decouple the relative position for each point from the global pose of the entire point cloud. Therefore, we extend current message passing networks with a module that learns orientations for each point. Before aggregating information from the neighbors of a point, the networks transforms the neighbors' coordinates based on the point's learned orientations. We provide formal proofs to show the equivariance of the proposed framework. Empirically, we demonstrate that our proposed method is competitive on both point cloud analysis and physical modeling tasks. Code is available at https://github.com/luost26/Equivariant-OrientedMP.

****

When Does Contrastive Visual Representation Learning Work?
Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, Serge Belongie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14755-14764

Recent self-supervised representation learning techniques have largely closed the gap between supervised and unsupervised learning on ImageNet classification. While the particulars of pretraining on ImageNet are now relatively well understood, the field still lacks widely accepted best practices for replicating this success on other datasets. As a first step in this direction, we study contrastive self-supervised learning on four diverse large-scale datasets. By looking through the lenses of data quantity, data domain, data quality, and task granularity, we provide new insights into the necessary conditions for successful self-supervised learning. Our key findings include observations such as: (i) the benefit of additional pretraining data beyond 500k images is modest, (ii) adding pretraining images from another domain does not lead to more general representations, (iii) corrupted pretraining images have a disparate impact on supervised and self-supervised pretraining, and (iv) contrastive learning lags far behind supervised learning on fine-grained visual classification tasks.

****

One Step at a Time: Long-Horizon Vision-and-Language Navigation With Milestones
Chan Hee Song, Jihyung Kil, Tai-Yu Pan, Brian M. Sadler, Wei-Lun Chao, Yu Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15482-15491

We study the problem of developing autonomous agents that can follow human instructions to infer and perform a sequence of actions to complete the underlying task. Significant progress has been made in recent years, especially for tasks with short horizons. However, when it comes to long-horizon tasks with extended sequences of actions, an agent can easily ignore some instructions or get stuck in the middle of the long instructions and eventually fail the task. To address this challenge, we propose a model-agnostic milestone-based task tracker (M-Track) to guide the agent and monitor its progress. Specifically, we propose a milestone builder that tags the instructions with navigation and interaction milestones which the agent needs to complete step by step, and a milestone checker that systemically checks the agent's progress in its current milestone and determines when to proceed to the next. On the challenging ALFRED dataset, our M-Track leads to a notable 33% and 52% relative improvement in unseen success rate over two competitive base models.

****

Node Representation Learning in Graph via Node-to-Neighbourhood Mutual Information Maximization

Wei Dong, Junsheng Wu, Yi Luo, Zongyuan Ge, Peng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16620-16629

The key towards learning informative node representations in graphs lies in how to gain contextual information from the neighbourhood. In this work, we present a simple-yet-effective self-supervised node representation learning strategy via directly maximizing the mutual information between the hidden representations of nodes and their neighbourhood, which can be theoretically justified by its link to graph smoothing. Following InfoNCE, our framework is optimized via a surrogate contrastive loss, where the positive selection underpins the quality and efficiency of representation learning. To this end, we propose a topology-aware positive sampling strategy, which samples positives from the neighbourhood by considering the structural dependencies between nodes and thus enables positive selection upfront. In the extreme case when only one positive is sampled, we fully avoid expensive neighbourhood aggregation. Our methods achieve promising performance on various node classification datasets. It is also worth mentioning by applying our loss function to MLP based node encoders, our methods can be orders of faster than existing solutions. Our codes and supplementary materials are available at https://github.com/dongwei156/n2n.

****

Point Cloud Pre-Training With Natural 3D Structures

Ryosuke Yamada, Hirokatsu Kataoka, Naoya Chiba, Yukiyasu Domae, Tetsuya Ogata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21283-21293

The construction of 3D point cloud datasets requires a great deal of human effort. Therefore, constructing a largescale 3D point clouds dataset is difficult. In order to remedy this issue, we propose a newly developed point cloud fractal database (PC-FractalDB), which is a novel family of formula-driven supervised learning inspired by fractal geometry encountered in natural 3D structures. Our research is based on the hypothesis that we could learn representations from more real-world 3D patterns than conventional 3D datasets by learning fractal geometry. We show how the PC-FractalDB facilitates solving several recent dataset-related problems in 3D scene understanding, such as 3D model collection and labor-intensive annotation. The experimental section shows how we achieved the performance rate of up to 61.9% and 59.0% for the ScanNetV2 and SUN RGB-D datasets, respectively, over the current highest scores obtained with the PointContrast, contrastive scene contexts (CSC), and RandomRooms. Moreover, the PC-FractalDB pre-trained model is especially effective in training with limited data. For example, in 10% of training data on ScanNetV2, the PC-FractalDB pre-trained VoteNet performs at 38.3%, which is +14.8% higher accuracy than CSC. Of particular note, we found that the proposed method achieves the highest results for 3D object detection pre-training in limited point cloud data.

****

Scene Consistency Representation Learning for Video Scene Segmentation
Haoqian Wu, Keyu Chen, Yanan Luo, Ruizhi Qiao, Bo Ren, Haozhe Liu, Weicheng Xie, Linlin Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14021-14030
A long-term video, such as a movie or TV show, is composed of various scenes, each of which represents a series of shots sharing the same semantic story. Spotting the correct scene boundary from the long-term video is a challenging task, since a model must understand the storyline of the video to figure out where a scene starts and ends. To this end, we propose an effective Self-Supervised Learning (SSL) framework to learn better shot representations from unlabeled long-term videos. More specifically, we present an SSL scheme to achieve scene consistency, while exploring considerable data augmentation and shuffling methods to boost the model generalizability. Instead of explicitly learning the scene boundary features as in the previous methods, we introduce a vanilla temporal model with less inductive bias to verify the quality of the shot features. Our method achieves the state-of-the-art performance on the task of Video Scene Segmentation. Additionally, we suggest a more fair and reasonable benchmark to evaluate the performance of Video Scene Segmentation methods. The code is made available.
****

Two Coupled Rejection Metrics Can Tell Adversarial Examples Apart
Tianyu Pang, Huishuai Zhang, Di He, Yinpeng Dong, Hang Su, Wei Chen, Jun Zhu, Tie-Yan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15223-15233
Correctly classifying adversarial examples is an essential but challenging requirement for safely deploying machine learning models. As reported in RobustBench, even the state-of-the-art adversarially trained models struggle to exceed 67% robust test accuracy on CIFAR-10, which is far from practical. A complementary way towards robustness is to introduce a rejection option, allowing the model to not return predictions on uncertain inputs, where confidence is a commonly used certainty proxy. Along with this routine, we find that confidence and a rectified confidence (R-Con) can form two coupled rejection metrics, which could provably distinguish wrongly classified inputs from correctly classified ones. This intriguing property sheds light on using coupling strategies to better detect and reject adversarial examples. We evaluate our rectified rejection (RR) module on CIFAR-10, CIFAR-10-C, and CIFAR-100 under several attacks including adaptive ones, and demonstrate that the RR module is compatible with different adversarial training frameworks on improving robustness, with little extra computation.
****

Exploiting Explainable Metrics for Augmented SGD
Mahdi S. Hosseini, Mathieu Tuli, Konstantinos N. Plataniotis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10296-10306
Explaining the generalization characteristics of deep learning is an emerging topic in advanced machine learning. There are several unanswered questions about how learning under stochastic optimization really works and why certain strategies are better than others. In this paper, we address the following question: can we probe intermediate layers of a deep neural network to identify and quantify the learning quality of each layer? With this question in mind, we propose new explainability metrics that measure the redundant information in a network's layers using a low-rank factorization framework and quantify a complexity measure that is highly correlated with the generalization performance of a given optimizer, network, and dataset. We subsequently exploit these metrics to augment the Stochastic Gradient Descent (SGD) optimizer by adaptively adjusting the learning rate in each layer to improve in generalization performance. Our augmented SGD -- dubbed RMSGD -- introduces minimal computational overhead compared to SOTA methods and outperforms them by exhibiting strong generalization characteristics across application, architecture, and dataset.
****

Semi-Supervised Video Semantic Segmentation With Inter-Frame Feature Reconstruction

Jiafan Zhuang, Zilei Wang, Yuan Gao; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2022, pp. 3263-3271
One major challenge for semantic segmentation in real-world scenarios is only li
mited pixel-level labels available due to high expense of human labor though a v
ast volume of video data is provided. Existing semi-supervised methods attempt t
o exploit unlabeled data in model training, but they just regard video as a set
of independent images. To better explore semi-supervised segmentation problem wi
th video data, we formulate a semi-supervised video semantic segmentation task i
n this paper. For this task, we observe that the overfitting is surprisingly sev
ere between labeled and unlabeled frames within a training video although they a
re very similar in style and contents. This is called inner-video overfitting, a
nd it would actually lead to inferior performance. To tackle this issue, we prop
ose a novel inter-frame feature reconstruction (IFR) technique to leverage the g
round-truth labels to supervise the model training on unlabeled frames. IFR is e
ssentially to utilize the internal relevance of different frames within a video.
 During training, IFR would enforce the feature distributions between labeled an
d unlabeled frames to be narrowed. Consequently, the inner-video overfitting iss
ue can be effectively alleviated. We conduct extensive experiments on Cityscapes
 and CamVid, and the results demonstrate the superiority of our proposed method
to previous state-of-the-art methods. The code is available at https://github.co
m/jfzhuang/IFR.
****

GenDR: A Generalized Differentiable Renderer
Felix Petersen, Bastian Goldluecke, Christian Borgelt, Oliver Deussen; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 4002-4011
In this work, we present and study a generalized family of differentiable render
ers. We discuss from scratch which components are necessary for differentiable r
endering and formalize the requirements for each component.We instantiate our ge
neral differentiable renderer, which generalizes existing differentiable rendere
rs like SoftRas and DIB-R, with an array of different smoothing distributions to
 cover a large spectrum of reasonable settings. We evaluate an array of differen
tiable renderer instantiations on the popular ShapeNet 3D reconstruction benchma
rk and analyze the implications of our results. Surprisingly, the simple uniform
 distribution yields the best overall results when averaged over 13 classes; in
general, however, the optimal choice of distribution heavily depends on the task
.
****

Improving Neural Implicit Surfaces Geometry With Patch Warping
François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, Mathieu
Aubry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2022, pp. 6260-6269
Neural implicit surfaces have become an important technique for multi-view 3D re
construction but their accuracy remains limited. In this paper, we argue that th
is comes from the difficulty to learn and render high frequency textures with ne
ural networks. We thus propose to add to the standard neural rendering optimizat
ion a direct photo-consistency term across the different views. Intuitively, we
optimize the implicit geometry so that it warps views on each other in a consist
ent way. We demonstrate that two elements are key to the success of such an appr
oach: (i) warping entire patches, using the predicted occupancy and normals of t
he 3D points along each ray, and measuring their similarity with a robust struct
ural similarity (SSIM); (ii) handling visibility and occlusion in such a way tha
t incorrect warps are not given too much importance while encouraging a reconstr
uction as complete as possible. We evaluate our approach, dubbed NeuralWarp, on
the standard DTU and EPFL benchmarks and show it outperforms state of the art un
supervised implicit surfaces reconstructions by over 20% on both datasets.
****

XYLayoutLM: Towards Layout-Aware Multimodal Networks for Visually-Rich Document
Understanding
Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, Liqing Zh

ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2022, pp. 4583-4592
Recently, various multimodal networks for Visually-Rich Document Understanding(V RDU) have been proposed, showing the promotion of transformers by integrating vi sual and layout information with the text embeddings. However, most existing app roaches utilize the position embeddings to incorporate the sequence information, neglecting the noisy improper reading order obtained by OCR tools. In this pape r, we propose a robust layout-aware multimodal network named XYLayoutLM to captu re and leverage rich layout information from proper reading orders produced by o ur Augmented XY Cut. Moreover, a Dilated Conditional Position Encoding module is proposed to deal with the input sequence of variable lengths, and it additional ly extracts local layout information from both textual and visual modalities whi le generating position embeddings. Experiment results show that our XYLayoutLM a chieves competitive results on document understanding tasks.
****

Amodal Segmentation Through Out-of-Task and Out-of-Distribution Generalization W ith a Bayesian Model
Yihong Sun, Adam Kortylewski, Alan Yuille; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1215-1224
Amodal completion is a visual task that humans perform easily but which is diffi cult for computer vision algorithms. The aim is to segment those object boundari es which are occluded and hence invisible. This task is particularly challenging for deep neural networks because data is difficult to obtain and annotate. Ther efore, we formulate amodal segmentation as an out-of-task and out-of-distributio n generalization problem. Specifically, we replace the fully connected classifie r in neural networks with a Bayesian generative model of the neural network feat ures. The model is trained from non-occluded images using bounding box annotatio ns and class labels only, but is applied to generalize out-of-task to object seg mentation and to generalize out-of-distribution to segment occluded objects. We demonstrate how such Bayesian models can naturally generalize beyond the trainin g task labels when they learn a prior that models the object's background contex t and shape. Moreover, by leveraging an outlier process, Bayesian models can fur ther generalize out-of-distribution to segment partially occluded objects and to predict their amodal object boundaries. Our algorithm outperforms alternative m ethods that use the same supervision by a large margin, and even outperforms met hods where annotated amodal segmentations are used during training, when the amo unt of occlusion is large. Code is publically available at https://github.com/an onymous-submission-vision/Amodal-Bayesian.
****

How Well Do Sparse ImageNet Models Transfer?
Eugenia Iofinova, Alexandra Peste, Mark Kurtz, Dan Alistarh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12266-12276
Transfer learning is a classic paradigm by which models pretrained on large "ups tream" datasets are adapted to yield good results on "downstream" specialized da tasets. Generally, more accurate models on the "upstream" dataset tend to provid e better transfer accuracy "downstream". In this work, we perform an in-depth in vestigation of this phenomenon in the context of convolutional neural networks (CNNs) trained on the ImageNet dataset, which have been pruned--that is, compress ed by sparsifiying their connections. We consider transfer using unstructured pr uned models obtained by applying several state-of-the-art pruning methods, inclu ding magnitude-based, second-order, re-growth, lottery-ticket, and regularizatio n approaches, in the context of twelve standard transfer tasks. In a nutshell, o ur study shows that sparse models can match or even outperform the transfer perf ormance of dense models, even at high sparsities, and, while doing so, can lead to significant inference and even training speedups. At the same time, we observ e and analyze significant differences in the behaviour of different pruning meth ods. The code is available at: https://github.com/IST-DASLab/sparse-imagenet-tra nsfer.
****

REX: Reasoning-Aware and Grounded Explanation
Shi Chen, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15586-15595
Effectiveness and interpretability are two essential properties for trustworthy AI systems. Most recent studies in visual reasoning are dedicated to improving the accuracy of predicted answers, and less attention is paid to explaining the rationales behind the decisions. As a result, they commonly take advantage of spurious biases instead of actually reasoning on the visual-textual data, and have yet developed the capability to explain their decision making by considering key information from both modalities. This paper aims to close the gap from three distinct perspectives: first, we define a new type of multi-modal explanations that explain the decisions by progressively traversing the reasoning process and grounding keywords in the images. We develop a functional program to sequentially execute different reasoning steps and construct a new dataset with 1,040,830 multi-modal explanations. Second, we identify the critical need to tightly couple important components across the visual and textual modalities for explaining the decisions, and propose a novel explanation generation method that explicitly models the pairwise correspondence between words and regions of interest. It improves the visual grounding capability by a considerable margin, resulting in enhanced interpretability and reasoning performance. Finally, with our new data and method, we perform extensive analyses to study the effectiveness of our explanation under different settings, including multi-task learning and transfer learning. Our code and data are available at https://github.com/szzexpoi/rex
****

Dynamic Dual-Output Diffusion Models
Yaniv Benny, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11482-11491
Iterative denoising-based generation, also known as denoising diffusion models, has recently been shown to be comparable in quality to other classes of generative models, and even surpass them. Including, in particular, Generative Adversarial Networks, which are currently the state of the art in many sub-tasks of image generation. However, a major drawback of this method is that it requires hundreds of iterations to produce a competitive result. Recent works have proposed solutions that allow for faster generation with fewer iterations, but the image quality gradually deteriorates with increasingly fewer iterations being applied during generation. In this paper, we reveal some of the causes that affect the generation quality of diffusion models, especially when sampling with few iterations, and come up with a simple, yet effective, solution to mitigate them. We consider two opposite equations for the iterative denoising, the first predicts the applied noise, and the second predicts the image directly. Our solution takes the two options and learns to dynamically alternate between them through the denoising process. Our proposed solution is general and can be applied to any existing diffusion model. As we show, when applied to various SOTA architectures, our solution immediately improves their generation quality, with negligible added complexity and parameters. We experiment on multiple datasets and configurations and run an extensive ablation study to support these findings.
****

StyleT2I: Toward Compositional and High-Fidelity Text-to-Image Synthesis
Zhiheng Li, Martin Renqiang Min, Kai Li, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18197-18207
Although progress has been made for text-to-image synthesis, previous methods fall short of generalizing to unseen or underrepresented attribute compositions in the input text. Lacking compositionality could have severe implications for robustness and fairness, e.g., inability to synthesize the face images of underrepresented demographic groups. In this paper, we introduce a new framework, StyleT2I, to improve the compositionality of text-to-image synthesis. Specifically, we propose a CLIP-guided Contrastive Loss to better distinguish different compositions among different sentences. To further improve the compositionality, we design a novel Semantic Matching Loss and a Spatial Constraint to identify attributes

' latent directions for intended spatial region manipulations, leading to better disentangled latent representations of attributes. Based on the identified latent directions of attributes, we propose Compositional Attribute Adjustment to adjust the latent code, resulting in better compositionality of image synthesis. In addition, we leverage the l_2-norm regularization of identified latent directions (norm penalty) to strike a nice balance between image-text alignment and image fidelity. In the experiments, we devise a new dataset split and an evaluation metric to evaluate the compositionality of text-to-image synthesis models. The results show that StyleT2I outperforms previous approaches in terms of the consistency between the input text and synthesized images and achieves higher fidelity.

****

JoinABLe: Learning Bottom-Up Assembly of Parametric CAD Joints

Karl D.D. Willis, Pradeep Kumar Jayaraman, Hang Chu, Yunsheng Tian, Yifei Li, Daniele Grandi, Aditya Sanghi, Linh Tran, Joseph G. Lambourne, Armando Solar-Lezama, Wojciech Matusik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15849-15860

Physical products are often complex assemblies combining a multitude of 3D parts modeled in computer-aided design (CAD) software. CAD designers build up these assemblies by aligning individual parts to one another using constraints called joints. In this paper we introduce JoinABLe, a learning-based method that assembles parts together to form joints. JoinABLe uses the weak supervision available in standard parametric CAD files without the help of object class labels or human guidance. Our results show that by making network predictions over a graph representation of solid models we can outperform multiple baseline methods with an accuracy (79.53%) that approaches human performance (80%). Finally, to support future research we release the AssemblyJoint dataset, containing assemblies with rich information on joints, contact surfaces, holes, and the underlying assembly graph structure.

****

CaDeX: Learning Canonical Deformation Coordinate Space for Dynamic Surface Representation via Neural Homeomorphism

Jiahui Lei, Kostas Daniilidis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6624-6634

While neural representations for static 3D shapes are widely studied, representations for deformable surfaces are limited to be template-dependent or to lack efficiency. We introduce Canonical Deformation Coordinate Space (CaDeX), a unified representation of both shape and nonrigid motion. Our key insight is the factorization of the deformation between frames by continuous bijective canonical maps (homeomorphisms) and their inverses that go through a learned canonical shape. Our novel deformation representation and its implementation are simple, efficient, and guarantee cycle consistency, topology preservation, and, if needed, volume conservation. Our modelling of the learned canonical shapes provides a flexible and stable space for shape prior learning. We demonstrate state-of-the-art performance in modelling a wide range of deformable geometries: human bodies, animal bodies, and articulated objects.

****

Canonical Voting: Towards Robust Oriented Bounding Box Detection in 3D Scenes

Yang You, Zelin Ye, Yujing Lou, Chengkun Li, Yong-Lu Li, Lizhuang Ma, Weiming Wang, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1193-1202

3D object detection has attracted much attention thanks to the advances in sensors and deep learning methods for point clouds. Current state-of-the-art methods like VoteNet regress direct offset towards object centers and box orientations with an additional Multi-Layer-Perceptron network. Both their offset and orientation predictions are not accurate due to the fundamental difficulty in rotation classification. In the work, we disentangle the direct offset into Local Canonical Coordinates (LCC), box scales and box orientations. Only LCC and box scales are regressed, while box orientations are generated by a canonical voting scheme. Finally, an LCC-aware back-projection checking algorithm iteratively cuts out bo

unding boxes from the generated vote maps, with the elimination of false positiv
es. Our model achieves state-of-the-art performance on three standard real-world
 benchmarks: ScanNet, SceneNN and SUN RGB-D. Our code is available on https://gi
thub.com/qq456cvb/CanonicalVoting.
****

V-Doc: Visual Questions Answers With Documents

Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyun
suk Chung, Soyeon Caren Han; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2022, pp. 21492-21498

We propose V-Doc, a question-answering tool using document images and PDF, mainl
y for researchers and general non-deep learning experts looking to generate, pro
cess, and understand the document visual question answering tasks. The V-Doc sup
ports generating and using both extractive and abstractive question-answer pairs
 using documents images. The extractive QA selects a subset of tokens or phrases
 from the document contents to predict the answers, while the abstractive QA rec
ognises the language in the content and generates the answer based on the traine
d model. Both aspects are crucial to understanding the documents, especially in
an image format. We include a detailed scenario of question generation for the a
bstractive QA task. V-Doc supports a wide range of datasets and models, and is h
ighly extensible through a declarative, framework-agnostic platform.
****

AEGNN: Asynchronous Event-Based Graph Neural Networks

Simon Schaefer, Daniel Gehrig, Davide Scaramuzza; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12371-1238
1

The best performing learning algorithms devised for event cameras work by first
converting events into dense representations that are then processed using stand
ard CNNs. However, these steps discard both the sparsity and high temporal resol
ution of events, leading to high computational burden and latency. For this reas
on, recent works have adopted Graph Neural Networks (GNNs), which process events
 as "static" spatio-temporal graphs, which are inherently "sparse". We take this
 trend one step further by introducing Asynchronous, Event-based Graph Neural Ne
tworks (AEGNNs), a novel event-processing paradigm that generalizes standard GNN
s to process events as "evolving" spatio-temporal graphs. AEGNNs follow efficien
t update rules that restrict recomputation of network activations only to the no
des affected by each new event, thereby significantly reducing both computation
and latency for event-by-event processing. AEGNNs are easily trained on synchron
ous inputs and can be converted to efficient, "asynchronous" networks at test ti
me. We thoroughly validate our method on object classification and detection tas
ks, where we show an up to a 200-fold reduction in computational complexity (FLO
Ps), with similar or even better performance than state-of-the-art asynchronous
methods. This reduction in computation directly translates to an 8-fold reductio
n in computational latency when compared to standard GNNs, which opens the door
to low-latency event-based processing.
****

Layer-Wised Model Aggregation for Personalized Federated Learning

Xiaosong Ma, Jie Zhang, Song Guo, Wenchao Xu; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10092-10101

Personalized Federated Learning (pFL) not only can capture the common priors fro
m broad range of distributed data, but also support customized models for hetero
geneous clients. Researches over the past few years have applied the weighted ag
gregation manner to produce personalized models, where the weights are determine
d by calibrating the distance of the entire model parameters or loss values, and
 have yet to consider the layer-level impacts to the aggregation process, leadin
g to lagged model convergence and inadequate personalization over non-IID datase
ts. In this paper, we propose a novel pFL training framework dubbed Layer-wised
Personalized Federated learning (pFedLA) that can discern the importance of each
 layer from different clients, and thus is able to optimize the personalized mod
el aggregation for clients with heterogeneous data. Specifically, we employ a de
dicated hypernetwork per client on the server side, which is trained to identify

the mutual contribution factors at layer granularity. Meanwhile, a parameterized mechanism is introduced to update the layer-wised aggregation weights to progressively exploit the inter-user similarity and realize accurate model personalization. Extensive experiments are conducted over different models and learning tasks, and we show that the proposed methods achieve significantly higher performance than state-of-the-art pFL methods.

****

Polarity Sampling: Quality and Diversity Control of Pre-Trained Generative Networks via Singular Values

Ahmed Imtiaz Humayun, Randall Balestriero, Richard Baraniuk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10641-10650

We present Polarity Sampling, a theoretically justified plug-and-play method for controlling the generation quality and diversity of any pre-trained deep generative network (DGN). Leveraging the fact that DGNs are, or can be approximated by, continuous piecewise affine splines, we derive the analytical DGN output space distribution as a function of the product of the DGN's Jacobian singular values raised to a power rho. We dub rho the polarity parameter and prove that rho focuses the DGN sampling on the modes (rho < 0) or anti-modes (rho > 0) of the DGN output space probability distribution. We demonstrate that nonzero polarity values achieve a better precision-recall (quality-diversity) Pareto frontier than standard methods, such as truncation, for a number of state-of-the-art DGNs. We also present quantitative and qualitative results on the improvement of overall generation quality (e.g., in terms of the Frechet Inception Distance) for a number of state-of-the-art DGNs, including StyleGAN3, BigGAN-deep, NVAE, for different conditional and unconditional image generation tasks. In particular, Polarity Sampling redefines the state-of-the-art for StyleGAN2 on the FFHQ Dataset to FID 2.57, StyleGAN2 on the LSUN Car Dataset to FID 2.27 and StyleGAN3 on the AFHQv2 Dataset to FID 3.95. Colab Demo: bit.ly/polarity-samp

****

Style-Structure Disentangled Features and Normalizing Flows for Diverse Icon Colorization

Yuan-kui Li, Yun-Hsuan Lien, Yu-Shuen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11244-11253

In this study, we present a colorization network that generates flat-color icons according to given sketches and semantic colorization styles. Specifically, our network contains a style-structure disentangled colorization module and a normalizing flow. The colorization module transforms a paired sketch image and style image into a flat-color icon. To enhance network generalization and the quality of icons, we present a pixel-wise decoder, a global style code, and a contour loss to reduce color gradients at flat regions and increase color discontinuity at boundaries. The normalizing flow maps Gaussian vectors to diverse style codes conditioned on the given semantic colorization label. This conditional sampling enables users to control attributes and obtain diverse colorization results. Compared to previous colorization methods built upon conditional generative adversarial networks, our approach enjoys the advantages of both high image quality and diversity. To evaluate its effectiveness, we compared the flat-color icons generated by our approach and recent colorization and image-to-image translation methods on various conditions. Experiment results verify that our method outperforms state-of-the-arts qualitatively and quantitatively.

****

Object-Aware Video-Language Pre-Training for Retrieval

Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3313-3322

Recently, by introducing large-scale dataset and strong transformer network, video-language pre-training has shown great success especially for retrieval. Yet, existing video-language transformer models do not explicitly fine-grained semantic align. In this work, we present Object-aware Transformers, an object-centric approach that extends video-language transformer to incorporate object represent

ations. The key idea is to leverage the bounding boxes and object tags to guide the training process. We evaluate our model on three standard sub-tasks of video-text matching on four widely used benchmarks. We also provide deep analysis and detailed ablation about the proposed method. We show clear improvement in performance across all tasks and datasets considered, demonstrating the value of a model that incorporates object representations into a video-language architecture. The code has been released in https://github.com/FingerRec/OA-Transformer.
****

OSKDet: Orientation-Sensitive Keypoint Localization for Rotated Object Detection
Dongchen Lu, Dongmei Li, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1182-1192
Rotated object detection is a challenging issue in computer vision field. Inadequate rotated representation and the confusion of parametric regression have been the bottleneck for high performance rotated detection. In this paper, we propose an orientation-sensitive keypoint based rotated detector OSKDet. First, we adopt a set of keypoints to represent the target and predict the keypoint heatmap on ROI to get the rotated box. By proposing the orientation-sensitive heatmap, OSKDet could learn the shape and direction of rotated target implicitly and has stronger modeling capabilities for rotated representation, which improves the localization accuracy and acquires high quality detection results. Second, we explore a new unordered keypoint representation paradigm, which could avoid the confusion of keypoint regression caused by rule based ordering. Furthermore, we propose a localization quality uncertainty module to better predict the classification score by the distribution uncertainty of keypoints heatmap. Experimental results on several public benchmarks show the state-of-the-art performance of OSKDet. Specifically, we achieve an AP of 80.91% on DOTA, 89.98% on HRSC2016, 97.27% on UCAS-AOD, and a F-measure of 92.18% on ICDAR2015, 81.43% on ICDAR2017, respectively.
****

MAT: Mask-Aware Transformer for Large Hole Image Inpainting
Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10758-10768
Recent studies have shown the importance of modeling long-range interactions in the inpainting problem. To achieve this goal, existing approaches exploit either standalone attention techniques or transformers, but usually under a low resolution in consideration of computational cost. In this paper, we present a novel transformer-based model for large hole inpainting, which unifies the merits of transformers and convolutions to efficiently process high-resolution images. We carefully design each component of our framework to guarantee the high fidelity and diversity of recovered images. Specifically, we customize an inpainting-oriented transformer block, where the attention module aggregates non-local information only from partial valid tokens, indicated by a dynamic mask. Extensive experiments demonstrate the state-of-the-art performance of the new model on multiple benchmark datasets. Code is released at https://github.com/fenglinglwb/MAT.
****

Exploring Geometric Consistency for Monocular 3D Object Detection
Qing Lian, Botao Ye, Ruijia Xu, Weilong Yao, Tong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1685-1694
This paper investigates the geometric consistency for monocular 3D object detection, which suffers from the ill-posed depth estimation. We first conduct a thorough analysis to reveal how existing methods fail to consistently localize objects when different geometric shifts occur. In particular, we design a series of geometric manipulations to diagnose existing detectors and then illustrate their vulnerability to consistently associate the depth with object apparent sizes and positions. To alleviate this issue, we propose four geometry-aware data augmentation approaches to enhance the geometric consistency of the detectors. We first modify some commonly used data augmentation methods for 2D images so that they can maintain geometric consistency in 3D spaces. We demonstrate such modification

s are important. In addition, we propose a 3D-specific image perturbation method that employs the camera movement. During the augmentation process, the camera system with the corresponding image is manipulated, while the geometric visual cues for depth recovery are preserved. We show that by using the geometric consistency constraints, the proposed augmentation techniques lead to improvements on the KITTI and nuScenes monocular 3D detection benchmarks with state-of-the-art results. In addition, we demonstrate that the augmentation methods are well suited for semi-supervised training and cross-dataset generalization.
****

Neural Window Fully-Connected CRFs for Monocular Depth Estimation
Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3916-3925
Estimating the accurate depth from a single image is challenging since it is inherently ambiguous and ill-posed. While recent works design increasingly complicated and powerful networks to directly regress the depth map, we take the path of CRFs optimization. Due to the expensive computation, CRFs are usually performed between neighborhoods rather than the whole graph. To leverage the potential of fully-connected CRFs, we split the input into windows and perform the FC-CRFs optimization within each window, which reduces the computation complexity and makes FC-CRFs feasible. To better capture the relationships between nodes in the graph, we exploit the multi-head attention mechanism to compute a multi-head potential function, which is fed to the networks to output an optimized depth map. Then we build a bottom-up-top-down structure, where this neural window FC-CRFs module serves as the decoder, and a vision transformer serves as the encoder. The experiments demonstrate that our method significantly improves the performance across all metrics on both the KITTI and NYUv2 datasets, compared to previous methods. Furthermore, the proposed method can be directly applied to panorama images and outperforms all previous panorama methods on the MatterPort3D dataset.
****

CodedVTR: Codebook-Based Sparse Voxel Transformer With Geometric Guidance
Tianchen Zhao, Niansong Zhang, Xuefei Ning, He Wang, Li Yi, Yu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1435-1444
Transformers have gained much attention by outperforming convolutional neural networks in many 2D vision tasks. However, they are known to have generalization problems and rely on massive-scale pre-training and sophisticated training techniques. When applying to 3D tasks, the irregular data structure and limited data scale add to the difficulty of transformer's application. We propose Codebook-based Voxel TRansformer), which improves data efficiency and generalization ability for 3D sparse voxel transformers. On the one hand, we propose the codebook-based attention that projects an attention space into its subspace represented by the combination of "prototypes" in a learnable codebook. It regularizes attention learning and improves generalization. On the other hand, we propose geometry-aware self-attention that utilizes geometric information (geometric pattern, density) to guide attention learning. CodedVTR could be embedded into existing sparse convolution-based methods, and bring consistent performance improvements for indoor and outdoor 3D semantic segmentation tasks.
****

Uncertainty-Aware Deep Multi-View Photometric Stereo
Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12601-12611
This paper presents a simple and effective solution to the longstanding classical multi-view photometric stereo (MVPS) problem. It is well-known that photometric stereo (PS) is excellent at recovering high-frequency surface details, whereas multi-view stereo (MVS) can help remove the low-frequency distortion due to PS and retain the global geometry of the shape. This paper proposes an approach that can effectively utilize such complementary strengths of PS and MVS. Our key idea is to combine them suitably while considering the per-pixel uncertainty of th

eir estimates. To this end, we estimate per-pixel surface normals and depth usin g an uncertainty-aware deep-PS network and deep-MVS network, respectively. Uncer tainty modeling helps select reliable surface normal and depth estimates at each pixel which then act as a true representative of the dense surface geometry. At each pixel, our approach either selects or discards deep-PS and deep-MVS networ k prediction depending on the prediction uncertainty measure. For dense, detaile d, and precise inference of the object's surface profile, we propose to learn th e implicit neural shape representation via a multilayer perceptron (MLP). Our ap proach encourages the MLP to converge to a natural zero-level set surface using the confident prediction from deep-PS and deep-MVS networks, providing superior dense surface reconstruction. Extensive experiments on the DiLiGenT-MV benchmark dataset show that our method provides high-quality shape recovery with a much l ower memory footprint while outperforming almost all of the existing approaches.
****

Coherent Point Drift Revisited for Non-Rigid Shape Matching and Registration
Aoxiang Fan, Jiayi Ma, Xin Tian, Xiaoguang Mei, Wei Liu; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 142 4-1434
In this paper, we explore a new type of extrinsic method to directly align two g eometric shapes with point-to-point correspondences in ambient space by recoveri ng a deformation, which allows more continuous and smooth maps to be obtained. S pecifically, the classic coherent point drift is revisited and generalizations h ave been proposed. First, by observing that the deformation model is essentially defined with respect to Euclidean space, we generalize the kernel method to non -Euclidean domains. This generally leads to better results for processing shapes , which are known as two-dimensional manifolds. Second, a generalized probabilis tic model is proposed to address the sensibility of coherent point drift method to local optima. Instead of directly optimizing over the objective of coherent p oint drift, the new model allows to focus on a group of most confident ones, thu s improves the robustness of the registration system. Experiments are conducted on multiple public datasets with comparison to state-of-the-art competitors, dem onstrating the superiority of our method which is both flexible and efficient to improve the matching accuracy due to our extrinsic alignment objective in ambie nt space.
****

Unleashing Potential of Unsupervised Pre-Training With Intra-Identity Regulariza tion for Person Re-Identification
Zizheng Yang, Xin Jin, Kecheng Zheng, Feng Zhao; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14298-14307
Existing person re-identification (ReID) methods typically directly load the pre -trained ImageNet weights for initialization. However, as a fine-grained classif ication task, ReID is more challenging and exists a large domain gap between Ima geNet classification. Inspired by the great success of self-supervised represent ation learning with contrastive objectives, in this paper, we design an Unsuperv ised Pre-training framework for ReID based on the contrastive learning (CL) pipe line, dubbed UP-ReID. During the pre-training, we attempt to address two critica l issues for learning fine-grained ReID features: (1) the augmentations in CL pi peline may distort the discriminative clues in person images. (2) the fine-grain ed local features of person images are not fully-explored. Therefore, we introdu ce an (I^2-)regularization in the UP-ReID, which is instantiated as two constrai nts coming from global image aspect and local patch aspect: a global consistency is enforced between augmented and original person images to increase robustness to augmentation, while an intrinsic contrastive constraint among local patches of each image is employed to fully explore the local discriminative clues. Exten sive experiments on multiple popular Re-ID datasets, including PersonX, Market15 01, CUHK03, and MSMT17, demonstrate that our UP-ReID pre-trained model can signi ficantly benefit the downstream ReID fine-tuning and achieve state-of-the-art pe rformance.
****

Align and Prompt: Video-and-Language Pre-Training With Entity Prompts

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, Steven C.H. Hoi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4953-4963

Video-and-language pre-training has shown promising improvements on various downstream tasks. Most previous methods capture cross-modal interactions with a transformer-based multimodal encoder, not fully addressing the misalignment between unimodal video and text features. Besides, learning fine-grained visual-language alignment usually requires off-the-shelf object detectors to provide object information, which is bottlenecked by the detector's limited vocabulary and expensive computation cost. We propose Align and Prompt: an efficient and effective video-and-language pre-training framework with better cross-modal alignment. First, we introduce a video-text contrastive (VTC) loss to align unimodal video-text features at the instance level, which eases the modeling of cross-modal interactions. Then, we propose a new visually-grounded pre-training task, prompting entity modeling (PEM), which aims to learn fine-grained region-entity alignment. To achieve this, we first introduce an entity prompter module, which is trained with VTC to produce the similarity between a video crop and text prompts instantiated with entity names. The PEM task then asks the model to predict the entity pseudo-labels (i.e normalized similarity scores) for randomly-selected video crops. The resulting pre-trained model achieves state-of-the-art performance on both text-video retrieval and videoQA, outperforming prior work by a substantial margin. Our code and pre-trained models are available at https://github.com/salesforce/ALPRO.

****

A Unified Query-Based Paradigm for Point Cloud Understanding

Zetong Yang, Li Jiang, Yanan Sun, Bernt Schiele, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8541-8551

3D point cloud understanding is an important component in autonomous driving and robotics. In this paper, we present a novel Embedding-Querying paradigm (EQ- Paradigm) for 3D understanding tasks including detection, segmentation and classification. EQ-Paradigm is a unified paradigm that enables combination of existing 3D backbone architectures with different task heads. Under the EQ- Paradigm, the input is first encoded in the embedding stage with an arbitrary feature extraction architecture, which is independent of tasks and heads. Then, the querying stage enables the encoded features for diverse task heads. This is achieved by introducing an intermediate representation, i.e., Q-representation, in the querying stage to bridge the embedding stage and task heads. We design a novel Q-Net as the querying stage network. Extensive experimental results on various 3D tasks show that EQ-Paradigm in tandem with Q-Net is a general and effective pipeline, which enables flexible collaboration of backbones and heads. It further boosts performance of state-of-the-art methods.

****

It's About Time: Analog Clock Reading in the Wild

Charig Yang, Weidi Xie, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2508-2517

In this paper, we present a framework for reading analog clocks in natural images or videos. Specifically, we make the following contributions: First, we create a scalable pipeline for generating synthetic clocks, significantly reducing the requirements for the labour-intensive annotations; Second, we introduce a clock recognition architecture based on spatial transformer networks (STN), which is trained end-to-end for clock alignment and recognition. We show that the model trained on the proposed synthetic dataset generalises towards real clocks with good accuracy, advocating a Sim2Real training regime; Third, to further reduce the gap between simulation and real data, we leverage the special property of "time", i.e.uniformity, to generate reliable pseudo-labels on real unlabelled clock videos, and show that training on these videos offers further improvements while still requiring zero manual annotations. Lastly, we introduce three benchmark datasets based on COCO, Open Images, and The Clock movie, with full annotations for time, accurate to the minute.

****
MSG-Transformer: Exchanging Local Spatial Information by Manipulating Messenger Tokens
Jiemin Fang, Lingxi Xie, Xinggang Wang, Xiaopeng Zhang, Wenyu Liu, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12063-12072
Transformers have offered a new methodology of designing neural networks for visual recognition. Compared to convolutional networks, Transformers enjoy the ability of referring to global features at each stage, yet the attention module brings higher computational overhead that obstructs the application of Transformers to process high-resolution visual data. This paper aims to alleviate the conflict between efficiency and flexibility, for which we propose a specialized token for each region that serves as a messenger (MSG). Hence, by manipulating these MSG tokens, one can flexibly exchange visual information across regions and the computational complexity is reduced. We then integrate the MSG token into a multi-scale architecture named MSG-Transformer. In standard image classification and object detection, MSG-Transformer achieves competitive performance and the inference on both GPU and CPU is accelerated. Code is available at https://github.com/hustvl/MSG-Transformer.
****
Cross Modal Retrieval With Querybank Normalisation
Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, Samuel Albanie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5194-5205
Profiting from large-scale training datasets, advances in neural architecture design and efficient inference, joint embeddings have become the dominant approach for tackling cross-modal retrieval. In this work we first show that, despite their effectiveness, state-of-the-art joint embeddings suffer significantly from the longstanding "hubness problem" in which a small number of gallery embeddings form the nearest neighbours of many queries. Drawing inspiration from the NLP literature, we formulate a simple but effective framework called Querybank Normalisation (QB-Norm) that re-normalises query similarities to account for hubs in the embedding space. QB-Norm improves retrieval performance without requiring retraining. Differently from prior work, we show that QB-Norm works effectively without concurrent access to any test set queries. Within the QB-Norm framework, we also propose a novel similarity normalisation method, the Dynamic Inverted Softmax, that is significantly more robust than existing approaches. We showcase QB-Norm across a range of cross modal retrieval models and benchmarks where it consistently enhances strong baselines beyond the state of the art. Code is available at https://vladbogo.github.io/QB-Norm/.
****
Contrastive Dual Gating: Learning Sparse Features With Contrastive Learning
Jian Meng, Li Yang, Jinwoo Shin, Deliang Fan, Jae-sun Seo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12257-12265
Contrastive learning (or its variants) has recently become a promising direction in the self-supervised learning domain, achieving similar performance as supervised learning with minimum fine-tuning. Despite the labeling efficiency, wide and large networks are required to achieve high accuracy, which incurs a high amount of computation and hinders the pragmatic merit of self-supervised learning. To effectively reduce the computation of insignificant features or channels, recent dynamic pruning algorithms for supervised learning employed auxiliary salience predictors. However, we found that such salience predictors cannot be easily trained when they are naively applied to contrastive learning from scratch. To address this issue, we propose contrastive dual gating(CDG), a novel dynamic pruning algorithm that skips the uninformative features during contrastive learning without hurting the trainability of the networks. We demonstrate the superiority of CDG with ResNet models for CIFAR-10, CIFAR-100, and ImageNet-100 datasets. Compared to our implementations of state-of-the-art dynamic pruning algorithms for self-supervised learning, CDG achieves up to 15% accuracy improvement for CIFAR

-10 dataset with higher computation reduction.
****

Universal Photometric Stereo Network Using Global Lighting Contexts
Satoshi Ikehata; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2022, pp. 12591-12600
This paper tackles a new photometric stereo task, named universal photometric st ereo. Unlike existing tasks that assumed specific physical lighting models; henc e, drastically limited their usability, a solution algorithm of this task is sup posed to work for objects with diverse shapes and materials under arbitrary ligh ting variations without assuming any specific models. To solve this extremely ch allenging task, we present a purely data-driven method, which eliminates the pri or assumption of lighting by replacing the recovery of physical lighting paramet ers with the extraction of the generic lighting representation, named global lig hting contexts. We use them like lighting parameters in a calibrated photometric stereo network to recover surface normal vectors pixelwisely. To adapt our netw ork to a wide variety of shapes, materials and lightings, it is trained on a new synthetic dataset which simulates the appearance of objects in the wild. Our me thod is compared with other state-of-the-art uncalibrated photometric stereo met hods on our test data to demonstrate the significance of our method.
****

Hire-MLP: Vision MLP via Hierarchical Rearrangement
Jianyuan Guo, Yehui Tang, Kai Han, Xinghao Chen, Han Wu, Chao Xu, Chang Xu, Yunh e Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 826-836
Previous vision MLPs such as MLP-Mixer and ResMLP accept linearly flattened imag e patches as input, making them inflexible for different input sizes and hard to capture spatial information. Such approach withholds MLPs from getting comparab le performance with their transformer-based counterparts and prevents them from becoming a general backbone for computer vision. This paper presents Hire-MLP, a simple yet competitive vision MLP architecture via Hierarchical rearrangement, which contains two levels of rearrangements. Specifically, the inner-region rear rangement is proposed to capture local information inside a spatial region, and the cross-region rearrangement is proposed to enable information communication b etween different regions and capture global context by circularly shifting all t okens along spatial directions. Extensive experiments demonstrate the effectiven ess of Hire-MLP as a versatile backbone for various vision tasks. In particular, Hire-MLP achieves competitive results on image classification, object detection and semantic segmentation tasks, e.g., 83.8% top-1 accuracy on ImageNet, 51.7% box AP and 44.8% mask AP on COCO val2017, and 49.9% mIoU on ADE20K, surpassing p revious transformer-based and MLP-based models with better trade-off for accurac y and throughput.
****

Ray3D: Ray-Based 3D Human Pose Estimation for Monocular Absolute 3D Localization
Yu Zhan, Fenghai Li, Renliang Weng, Wongun Choi; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13116-13125
In this paper, we propose a novel monocular ray-based 3D (Ray3D) absolute human pose estimation with calibrated camera. Accurate and generalizable absolute 3D h uman pose estimation from monocular 2D pose input is an ill-posed problem. To ad dress this challenge, we convert the input from pixel space to 3D normalized ray s. This conversion makes our approach robust to camera intrinsic parameter chang es. To deal with the in-the-wild camera extrinsic parameter variations, Ray3D ex plicitly takes the camera extrinsic parameters as an input and jointly models th e distribution between the 3D pose rays and camera extrinsic parameters. This no vel network design is the key to the outstanding generalizability of Ray3D appro ach. To have a comprehensive understanding of how the camera intrinsic and extri nsic parameter variations affect the accuracy of absolute 3D key-point localizat ion, we conduct in-depth systematic experiments on three single person 3D benchm arks as well as one synthetic benchmark. These experiments demonstrate that our method significantly outperforms existing state-of-the-art models. Our code and the synthetic dataset are available at https://github.com/YxZhxn/Ray3D.

****
Occluded Human Mesh Recovery
Rawal Khirodkar, Shashank Tripathi, Kris Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1715-1725
Top-down methods for monocular human mesh recovery have two stages: (1) detect human bounding boxes; (2) treat each bounding box as an independent single-human mesh recovery task. Unfortunately, the single-human assumption does not hold in images with multi-human occlusion and crowding. Consequently, top-down methods have difficulties in recovering accurate 3D human meshes under severe person-person occlusion. To address this, we present Occluded Human Mesh Recovery (OCHMR) - a novel top-down mesh recovery approach that incorporates image spatial context to overcome the limitations of the single-human assumption. The approach is conceptually simple and can be applied to any existing top-down architecture. Along with the input image, we condition the top-down model on spatial context from the image in the form of body-center heatmaps. To reason from the predicted body centermaps, we introduce Contextual Normalization (CoNorm) blocks to adaptively modulate intermediate features of the top-down model. The contextual conditioning helps our model disambiguate between two severely overlapping human bounding-boxes, making it robust to multi-person occlusion. Compared with state-of-the-art methods, OCHMR achieves superior performance on challenging multi-person benchmarks like 3DPW, CrowdPose, and OCHuman. Specifically, our proposed contextual reasoning architecture applied to the SPIN model with ResNet-50 backbone results in 75.2 PMPJPE on 3DPW-PC, 23.6 AP on CrowdPose, and 37.7 AP on OCHuman datasets, a significant improvement of 6.9 mm, 6.4 AP, and 20.8 AP respectively over the baseline. Code and models will be released.
****
Multi-Object Tracking Meets Moving UAV
Shuai Liu, Xin Li, Huchuan Lu, You He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8876-8885
Multi-object tracking in unmanned aerial vehicle (UAV) videos is an important vision task and can be applied in a wide range of applications. However, conventional multi-object trackers do not work well on UAV videos due to the challenging factors of irregular motion caused by moving camera and view change in 3D directions. In this paper, we propose a UAVMOT network specially for multi-object tracking in UAV views. The UAVMOT introduces an ID feature update module to enhance the object's feature association. To better handle the complex motions under UAV views, we develop an adaptive motion filter module. In addition, a gradient balanced focal loss is used to tackle the imbalance categories and small objects detection problem. Experimental results on the VisDrone2019 and UAVDT datasets demonstrate that the proposed UAVMOT achieves considerable improvement against the state-of-the-art tracking methods on UAV videos.
****
ASM-Loc: Action-Aware Segment Modeling for Weakly-Supervised Temporal Action Localization
Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13925-13935
Weakly-supervised temporal action localization aims to recognize and localize action segments in untrimmed videos given only video-level action labels for training. Without the boundary information of action segments, existing methods mostly rely on multiple instance learning (MIL), where the predictions of unlabeled instances (i.e., video snippets) are supervised by classifying labeled bags (i.e., untrimmed videos). However, this formulation typically treats snippets in a video as independent instances, ignoring the underlying temporal structures within and across action segments. To address this problem, we propose \system, a novel WTAL framework that enables explicit, action-aware segment modeling beyond standard MIL-based methods. Our framework entails three segment-centric components: (i) dynamic segment sampling for compensating the contribution of short actions; (ii) intra- and inter-segment attention for modeling action dynamics and capturing temporal dependencies; (iii) pseudo instance-level supervision for improvin

g action boundary prediction. Furthermore, a multi-step refinement strategy is p
roposed to progressively improve action proposals along the model training proce
ss. Extensive experiments on THUMOS-14 and ActivityNet-v1.3 demonstrate the effe
ctiveness of our approach, establishing new state of the art on both datasets. T
he code and models are publicly available at https://github.com/boheumd/ASM-Loc.
****

Uncertainty-Guided Probabilistic Transformer for Complex Action Recognition
Hongji Guo, Hanjing Wang, Qiang Ji; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 20052-20061
A complex action consists of a sequence of atomic actions that interact with eac
h other over a relatively long period of time. This paper introduces a probabili
stic model named Uncertainty-Guided Probabilistic Transformer (UGPT) for complex
 action recognition. The self-attention mechanism of a Transformer is used to ca
pture the complex and long-term dynamics of the complex actions. By explicitly m
odeling the distribution of the attention scores, we extend the deterministic Tr
ansformer to a probabilistic Transformer in order to quantify the uncertainty of
 the prediction. The model prediction uncertainty is used to improve both traini
ng and inference. Specifically, we propose a novel training strategy by introduc
ing a majority model and a minority model based on the epistemic uncertainty. Du
ring the inference, the prediction is jointly made by both models through a dyna
mic fusion strategy. Our method is validated on the benchmark datasets, includin
g Breakfast Actions, MultiTHUMOS, and Charades. The experiment results show that
 our model achieves the state-of-the-art performance under both sufficient and i
nsufficient data.
****

Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs
Xiaohan Ding, Xiangyu Zhang, Jungong Han, Guiguang Ding; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 119
63-11975
We revisit large kernel design in modern convolutional neural networks (CNNs). I
nspired by recent advances in vision transformers (ViTs), in this paper, we demo
nstrate that using a few large convolutional kernels instead of a stack of small
 kernels could be a more powerful paradigm. We suggested five guidelines, e.g.,
applying re-parameterized large depth-wise convolutions, to design efficient hig
h-performance large-kernel CNNs. Following the guidelines, we propose RepLKNet,
a pure CNN architecture whose kernel size is as large as 31x31, in contrast to c
ommonly used 3x3. RepLKNet greatly closes the performance gap between CNNs and V
iTs, e.g., achieving comparable or superior results than Swin Transformer on Ima
geNet and a few typical downstream tasks, with lower latency. RepLKNet also show
s nice scalability to big data and large models, obtaining 87.8% top-1 accuracy
on ImageNet and 56.0% mIoU on ADE20K, which is very competitive among the state-
of-the-arts with similar model sizes. Our study further reveals that, in contras
t to small-kernel CNNs, large-kernel CNNs have much larger effective receptive f
ields and higher shape bias rather than texture bias. Code & models at https://g
ithub.com/megvii-research/RepLKNet.
****

End-to-End Multi-Person Pose Estimation With Transformers
Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, Wenming Tan; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11069-1
1078
Current methods of multi-person pose estimation typically treat the localization
 and association of body joints separately. In this paper, we propose the first
fully end-to-end multi-person Pose Estimation framework with TRansformers, terme
d PETR. Our method views pose estimation as a hierarchical set prediction proble
m and effectively removes the need for many hand-crafted modules like RoI croppi
ng, NMS and grouping post-processing. In PETR, multiple pose queries are learned
 to directly reason a set of full-body poses. Then a joint decoder is utilized t
o further refine the poses by exploring the kinematic relations between body joi
nts. With the attention mechanism, the proposed method is able to adaptively att
end to the features most relevant to target keypoints, which largely overcomes t

he feature misalignment difficulty in pose estimation and improves the performan ce considerably. Extensive experiments on the MS COCO and CrowdPose benchmarks s how that PETR plays favorably against state-of-the-art approaches in terms of bo th accuracy and efficiency. The code and models are available at https://github. com/hikvision-research/opera.
****

REGTR: End-to-End Point Cloud Correspondences With Transformers
Zi Jian Yew, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2022, pp. 6677-6686
Despite recent success in incorporating learning into point cloud registration, many works focus on learning feature descriptors and continue to rely on nearest -neighbor feature matching and outlier filtering through RANSAC to obtain the fi nal set of correspondences for pose estimation. In this work, we conjecture that attention mechanisms can replace the role of explicit feature matching and RANS AC, and thus propose an end-to-end framework to directly predict the final set o f correspondences. We use a network architecture consisting primarily of transfo rmer layers containing self and cross attentions, and train it to predict the pr obability each point lies in the overlapping region and its corresponding positi on in the other point cloud. The required rigid transformation can then be estim ated directly from the predicted correspondences without further post-processing . Despite its simplicity, our approach achieves state-of-the-art performance on 3DMatch and ModelNet benchmarks. Our source code can be found at https://github. com/yewzijian/RegTR.
****

Neural 3D Scene Reconstruction With the Manhattan-World Assumption
Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, Xiao wei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5511-5520
This paper addresses the challenge of reconstructing 3D indoor scenes from multi -view images. Many previous works have shown impressive reconstruction results o n textured objects, but they still have difficulty in handling low-textured plan ar regions, which are common in indoor scenes. An approach to solving this issue is to incorporate planer constraints into the depth map estimation in multi-vie w stereo-based methods, but the per-view plane estimation and depth optimization lack both efficiency and multi-view consistency. In this work, we show that the planar constraints can be conveniently integrated into the recent implicit neur al representation-based reconstruction methods. Specifically, we use an MLP netw ork to represent the signed distance function as the scene geometry. Based on th e Manhattan-world assumption, planar constraints are employed to regularize the geometry in floor and wall regions predicted by a 2D semantic segmentation netwo rk. To resolve the inaccurate segmentation, we encode the semantics of 3D points with another MLP and design a novel loss that jointly optimizes the scene geome try and semantics in 3D space. Experiments on ScanNet and 7-Scenes datasets show that the proposed method outperforms previous methods by a large margin on 3D r econstruction quality. The code and supplementary materials are available at htt ps://zju3dv.github.io/manhattan_sdf.
****

V2C: Visual Voice Cloning
Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, Qi Wu; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 21242-21251
Existing Voice Cloning (VC) tasks aim to convert a paragraph text to a speech wi th desired voice specified by a reference audio. This has significantly boosted the development of artificial speech applications. However, there also exist man y scenarios that cannot be well reflected by these VC tasks, such as movie dubbi ng, which requires the speech to be with emotions consistent with the movie plot s. To fill this gap, in this work we propose a new task named Visual Voice Cloni ng (V2C), which seeks to convert a paragraph of text to a speech with both desir ed voice specified by a reference audio and desired emotion specified by a refer ence video. To facilitate research in this field, we construct a dataset, V2C-An

imation, and propose a strong baseline based on existing state-of-the-art (SoTA) VC techniques. Our dataset contains 10,217 animated movie clips covering a large variety of genres (e.g., Comedy, Fantasy) and emotions (e.g., happy, sad). We further design a set of evaluation metrics, named MCD-DTW-SL, which help evaluate the similarity between ground-truth speeches and the synthesised ones. Extensive experimental results show that even SoTA VC methods cannot generate satisfying speeches for our V2C task. We hope the proposed new task together with the constructed dataset and evaluation metric will facilitate the research in the field of voice cloning and broader vision-and-language community. Source code and dataset will be released in https://github.com/chenqi008/V2C.
****

Revisiting AP Loss for Dense Object Detection: Adaptive Ranking Pair Selection
Dongli Xu, Jinhong Deng, Wen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14187-14196
Average precision (AP) loss has recently shown promising performance on the dense object detection task. However, a deep understanding of how AP loss affects the detector from a pairwise ranking perspective has not yet been developed. In this work, we revisit the average precision (AP) loss and reveal that the crucial element is that of selecting the ranking pairs between positive and negative samples. Based on this observation, we propose two strategies to improve the AP loss. The first of these is a novel Adaptive Pairwise Error (APE) loss that focusing on ranking pairs in both positive and negative samples. Moreover, we select more accurate ranking pairs by exploiting the normalized ranking scores and localization scores with a clustering algorithm. Experiments conducted on the MS-COCO dataset support our analysis and demonstrate the superiority of our proposed method compared with current classification and ranking loss. The code is available at https://github.com/Xudangliatiger/APE-Loss.
****

3DeformRS: Certifying Spatial Deformations on Point Clouds
Gabriel Pérez S., Juan C. Pérez, Motasem Alfarra, Silvio Giancola, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15169-15179
3D computer vision models are commonly used in security-critical applications such as autonomous driving and surgical robotics. Emerging concerns over the robustness of these models against real-world deformations must be addressed practically and reliably. In this work, we propose 3DeformRS, a method to certify the robustness of point cloud Deep Neural Networks (DNNs) against real-world deformations. We developed 3DeformRS by building upon recent work that generalized Randomized Smoothing (RS) from pixel-intensity perturbations to vector-field deformations. In particular, we specialized RS to certify DNNs against parameterized deformations (e.g. rotation, twisting), while enjoying practical computational costs. We leverage the virtues of 3DeformRS to conduct a comprehensive empirical study on the certified robustness of four representative point cloud DNNs on two datasets and against seven different deformations. Compared to previous approaches for certifying point cloud DNNs, 3DeformRS is fast, scales well with point cloud size, and provides comparable-to-better certificates. For instance, when certifying a plain PointNet against a 3deg z-rotation on 1024-point clouds, 3DeformRS grants a certificate 3x larger and 20x faster than previous work.
****

ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses
Bastian Wandt, James J. Little, Helge Rhodin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6635-6645
Human pose estimation from single images is a challenging problem that is typically solved by supervised learning. Unfortunately, labeled training data does not yet exist for many human activities since 3D annotation requires dedicated motion capture systems. Therefore, we propose an unsupervised approach that learns to predict a 3D human pose from a single image while only being trained with 2D pose data, which can be crowd-sourced and is already widely available. To this end, we estimate the 3D pose that is most likely over random projections, with the

likelihood estimated using normalizing flows on 2D poses. While previous work requires strong priors on camera rotations in the training data set, we learn the distribution of camera angles which significantly improves the performance. Another part of our contribution is to stabilize training with normalizing flows on high-dimensional 3D pose data by first projecting the 2D poses to a linear subspace. We outperform state-of-the-art in unsupervised human pose estimation on the benchmark dataset Human3.6M in all metrics.

****

MAD: A Scalable Dataset for Language Grounding in Videos From Movie Audio Descriptions

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5026-5035

The recent and increasing interest in video-language research has driven the development of large-scale datasets that enable data-intensive machine learning techniques. In comparison, limited effort has been made at assessing the fitness of these datasets for the video-language grounding task. Recent works have begun to discover significant limitations in these datasets, suggesting that state-of-the-art techniques commonly overfit to hidden dataset biases. In this work, we present MAD (Movie Audio Descriptions), a novel benchmark that departs from the paradigm of augmenting existing video datasets with text annotations and focuses on crawling and aligning available audio descriptions of mainstream movies. MAD contains over 384,000 natural language sentences grounded in over 1,200 hours of videos and exhibits a significant reduction in the currently diagnosed biases for video-language grounding datasets. MAD's collection strategy enables a novel and more challenging version of video-language grounding, where short temporal moments (typically seconds long) must be accurately grounded in diverse long-form videos that can last up to three hours. We have released MAD's data and baselines code at https://github.com/Soldelli/MAD.

****

EvUnroll: Neuromorphic Events Based Rolling Shutter Image Correction

Xinyu Zhou, Peiqi Duan, Yi Ma, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17775-17784

This paper proposes to use neuromorphic events for correcting rolling shutter (RS) images as consecutive global shutter (GS) frames. RS effect introduces edge distortion and region occlusion into images caused by row-wise readout of CMOS sensors. We introduce a novel computational imaging setup consisting of an RS sensor and an event sensor, and propose a neural network called EvUnroll to solve this problem by exploring the high-temporal-resolution property of events. We use events to bridge a spatio-temporal connection between RS and GS, establish a flow estimation module to correct edge distortions, and design a synthesis-based restoration module to restore occluded regions. The results of two branches are fused through a refining module to generate corrected GS images. We further propose datasets captured by a high-speed camera and an RS-Event hybrid camera system for training and testing our network. Experimental results on both public and proposed datasets show a systematic performance improvement compared to state-of-the-art methods.

****

Gait Recognition in the Wild With Dense 3D Representations and a Benchmark

Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20228-20237

Existing studies for gait recognition are dominated by 2D representations like the silhouette or skeleton of the human body in constrained scenes. However, humans live and walk in the unconstrained 3D space, so projecting the 3D human body onto the 2D plane will discard a lot of crucial information like the viewpoint, shape, and dynamics for gait recognition. Therefore, this paper aims to explore dense 3D representations for gait recognition in the wild, which is a practical yet neglected problem. In particular, we propose a novel framework to explore the 3D Skinned Multi-Person Linear (SMPL) model of the human body for gait recogni

tion, named SMPLGait. Our framework has two elaborately-designed branches of which one extracts appearance features from silhouettes, the other learns knowledge of 3D viewpoints and shapes from the 3D SMPL model. In addition, due to the lack of suitable datasets, we build the first large-scale 3D representation-based gait recognition dataset, named Gait3D. It contains 4,000 subjects and over 25,000 sequences extracted from 39 cameras in an unconstrained indoor scene. More importantly, it provides 3D SMPL models recovered from video frames which can provide dense 3D information of body shape, viewpoint, and dynamics. Based on Gait3D, we comprehensively compare our method with existing gait recognition approaches, which reflects the superior performance of our framework and the potential of 3D representations for gait recognition in the wild. The code and dataset are available at: https://gait3d.github.io.

****

ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis

Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2750-2760

Estimating the articulated 3D hand-object pose from a single RGB image is a highly ambiguous and challenging problem, requiring large-scale datasets that contain diverse hand poses, object types, and camera viewpoints. Most real-world datasets lack these diversities. In contrast, data synthesis can easily ensure those diversities separately. However, constructing both valid and diverse hand-object interactions and efficiently learning from the vast synthetic data is still challenging. To address the above issues, we propose ArtiBoost, a lightweight online data enhancement method. ArtiBoost can cover diverse hand-object poses and camera viewpoints through sampling in a Composited hand-object Configuration and Viewpoint space (CCV-space) and can adaptively enrich the current hard-discernable items by loss-feedback and sample re-weighting. ArtiBoost alternatively performs data exploration and synthesis within a learning pipeline, and those synthetic data are blended into real-world source data for training. We apply ArtiBoost on a simple learning baseline network and witness the performance boost on several hand-object benchmarks. Our models and code are available at https://github.com/lixiny/ArtiBoost.

****

Temporal Context Matters: Enhancing Single Image Prediction With Disease Progression Representations

Aishik Konwer, Xuan Xu, Joseph Bae, Chao Chen, Prateek Prasanna; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18824-18835

Clinical outcome or severity prediction from medical images has largely focused on learning representations from single-timepoint or snapshot scans. It has been shown that disease progression can be better characterized by temporal imaging. We therefore hypothesized that outcome predictions can be improved by utilizing the disease progression information from sequential images. We present a deep learning approach that leverages temporal progression information to improve clinical outcome predictions from single-timepoint images. In our method, a self-attention based Temporal Convolutional Network (TCN) is used to learn a representation that is most reflective of the disease trajectory. Meanwhile, a Vision Transformer is pretrained in a self-supervised fashion to extract features from single-timepoint images. The key contribution is to design a recalibration module that employs maximum mean discrepancy loss (MMD) to align distributions of the above two contextual representations. We train our system to predict clinical outcomes and severity grades from single-timepoint images. Experiments on chest and osteoarthritis radiography datasets demonstrate that our approach outperforms other state-of-the-art techniques.

****

QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection

Chenhongyi Yang, Zehao Huang, Naiyan Wang; Proceedings of the IEEE/CVF Conferenc

e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13668-13677
While general object detection with deep learning has achieved great success in the past few years, the performance and efficiency of detecting small objects are far from satisfactory. The most common and effective way to promote small object detection is to use high-resolution images or feature maps. However, both approaches induce costly computation since the computational cost grows squarely as the size of images and features increases. To get the best of two worlds, we propose QueryDet that uses a novel query mechanism to accelerate the inference speed of feature-pyramid based object detectors. The pipeline composes two steps: it first predicts the coarse locations of small objects on low-resolution features and then computes the accurate detection results using high-resolution features sparsely guided by those coarse positions. In this way, we can not only harvest the benefit of high-resolution feature maps but also avoid useless computation for the background area. On the popular COCO dataset, the proposed method improves the detection mAP by 1.0 and mAP small by 2.0, and the high-resolution inference speed is improved to 3.0x on average. On VisDrone dataset, which contains more small objects, we create a new state-of-the-art while gaining a 2.3x high-resolution acceleration on average. Code is available at https://github.com/ ChenhongyiYang/QueryDet-PyTorch.
****

IDEA-Net: Dynamic 3D Point Cloud Interpolation via Deep Embedding Alignment
Yiming Zeng, Yue Qian, Qijian Zhang, Junhui Hou, Yixuan Yuan, Ying He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6338-6347
This paper investigates the problem of temporally interpolating dynamic 3D point clouds with large non-rigid deformation. We formulate the problem as estimation of point-wise trajectories (i.e., smooth curves) and further reason that temporal irregularity and under-sampling are two major challenges. To tackle the challenges, we propose IDEA-Net, an end-to-end deep learning framework, which disentangles the problem under the assistance of the explicitly learned temporal consistency. Specifically, we propose a temporal consistency learning module to align two consecutive point cloud frames point-wisely, based on which we can employ linear interpolation to obtain coarse trajectories/in-between frames. To compensate the high-order nonlinear components of trajectories, we apply aligned feature embeddings that encode local geometry properties to regress point-wise increments, which are combined with the coarse estimations. We demonstrate the effectiveness of our method on various point cloud sequences and observe large improvement over state-of-the-art methods both quantitatively and visually. Our framework can bring benefits to 3D motion data acquisition. The source code is publicly available at https://github.com/ZENGYIMING-EAMON/IDEA-Net.git.
****

UniCon: Combating Label Noise Through Uniform Selection and Contrastive Learning
Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9676-9686
Supervised deep learning methods require a large repository of annotated data; hence, label noise is inevitable. Training with such noisy data negatively impacts the generalization performance of deep neural networks. To combat label noise, recent state-of-the-art methods employ some sort of sample selection mechanism to select a possibly clean subset of data. Next, an off-the-shelf semi-supervised learning method is used for training where rejected samples are treated as unlabeled data. Our comprehensive analysis shows that current selection methods disproportionately select samples from easy (fast learnable) classes while rejecting those from relatively harder ones. This creates class imbalance in the selected clean set and in turn, deteriorates performance under high label noise. In this work, we propose UNICON, a simple yet effective sample selection method which is robust to high label noise. To address the disproportionate selection of easy and hard samples, we introduce a Jensen-Shannon divergence based uniform selection mechanism which does not require any probabilistic modeling and hyperparameter tuning. We complement our selection method with contrastive learning to furth

er combat the memorization of noisy labels. Extensive experimentation on multipl e benchmark datasets demonstrates the effectiveness of UNICON; we obtain an 11.4 % improvement over the current state-of-the-art on CIFAR100 dataset with a 90% n oise rate. Our code is publicly available.
****

Learning From All Vehicles
Dian Chen, Philipp Krähenbühl; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 17222-17231
In this paper, we present a system to train driving policies from experiences co llected not just from the ego-vehicle, but all vehicles that it observes. This s ystem uses the behaviors of other agents to create more diverse driving scenario s without collecting additional data. The main difficulty in learning from other vehicles is that there is no sensor information. We use a set of supervisory ta sks to learn an intermediate representation that is invariant to the viewpoint o f the controlling vehicle. This not only provides a richer signal at training ti me but also allows more complex reasoning during inference. Learning how all veh icles drive helps predict their behavior at test time and can avoid collisions. We evaluate this system in closed-loop driving simulations. Our system outperfor ms all prior methods on the public CARLA Leaderboard by a wide margin, improving driving score by 25 and route completion rate by 24 points.
****

BEHAVE: Dataset and Method for Tracking Human Object Interactions
Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Chris tian Theobalt, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2022, pp. 15935-15946
Modelling interactions between humans and objects in natural environments is cen tral to many applications including gaming, virtual and mixed reality, as well a s human behavior analysis and human-robot collaboration. This challenging operat ion scenario requires generalization to vast number of objects, scenes, and huma n actions. Unfortunately, there exist no such dataset. Moreover, this data needs to be acquired in diverse natural environments, which rules out 4D scanners and marker based capture systems. We present BEHAVE dataset, the first full body hu man-object interaction dataset with multi-view RGBD frames and corresponding 3D SMPL and object fits along with the annotated contacts between them. We record 15k frames at 5 locations with 8 subjects performing a wide range of interaction s with 20 common objects. We use this data to learn a model that can jointly tra ck humans and objects in natural environments with an easy-to-use portable multi -camera setup. Our key insight is to predict correspondences from the human and the object to a statistical body model to obtain human-object contacts during in teractions. Our approach can record and track not just the humans and objects bu t also their interactions, modeled as surface contacts, in 3D. Our code and data can be found at: http://virtualhumans.mpi-inf.mpg.de/behave.
****

Disentangled3D: Learning a 3D Generative Model With Disentangled Geometry and Ap pearance From Monocular Images
Ayush Tewari, Mallikarjun B R, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christ ian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2022, pp. 1516-1525
Learning 3D generative models from a dataset of monocular images enables self-su pervised 3D reasoning and controllable synthesis. State-of-the-art 3D generative models are GANs which use neural 3D volumetric representations for synthesis. I mages are synthesized by rendering the volumes from a given camera. These models can disentangle the 3D scene from the camera viewpoint in any generated image. However, most models do not disentangle other factors of image formation, such a s geometry and appearance. In this paper, we design a 3D GAN which can learn a d isentangled model of objects, just from monocular observations. Our model can di sentangle the geometry and appearance variations in the scene, i.e., we can inde pendently sample from the geometry and appearance spaces of the generative model . This is achieved using a novel non-rigid deformable scene formulation. A 3D vo lume which represents an object instance is computed as a non-rigidly deformed c

anonical 3D volume. Our method learns the canonical volume, as well as its defor
mations, jointly during training. This formulation also helps us improve the dis
entanglement between the 3D scene and the camera viewpoints using a novel pose r
egularization loss defined on the 3D deformation field. In addition, we further
model the inverse deformations, enabling the computation of dense correspondence
s between images generated by our model. Finally, we design an approach to embed
 real images onto the latent space of our disentangled generative model, enablin
g editing of real images.
****

Revisiting Random Channel Pruning for Neural Network Compression
Yawei Li, Kamil Adamczewski, Wen Li, Shuhang Gu, Radu Timofte, Luc Van Gool; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2022, pp. 191-201
Channel (or 3D filter) pruning serves as an effective way to accelerate the infe
rence of neural networks. There has been a flurry of algorithms that try to solv
e this practical problem, each being claimed effective in some ways. Yet, a benc
hmark to compare those algorithms directly is lacking, mainly due to the complex
ity of the algorithms and some custom settings such as the particular network co
nfiguration or training procedure. A fair benchmark is important for the further
 development of channel pruning. Meanwhile, recent investigations reveal that th
e channel configurations discovered by pruning algorithms are at least as import
ant as the pre-trained weights. This gives channel pruning a new role, namely se
arching the optimal channel configuration. In this paper, we try to determine th
e channel configuration of the pruned models by random search. The proposed appr
oach provides a new way to compare different methods, namely how well they behav
e compared with random pruning. We show that this simple strategy works quite we
ll compared with other channel pruning methods. We also show that under this set
ting, there are surprisingly no clear winners among different channel importance
 evaluation methods, which then may tilt the research efforts into advanced chan
nel configuration searching methods. Code will be released at https://github.com
/ofsoundof/random_channel_pruning.
****

One-Bit Active Query With Contrastive Pairs
Yuhang Zhang, Xiaopeng Zhang, Lingxi Xie, Jie Li, Robert C. Qiu, Hengtong Hu, Qi
 Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2022, pp. 9697-9705
How to achieve better results with fewer labeling costs remains a challenging ta
sk. In this paper, we present a new active learning framework, which for the fir
st time incorporates contrastive learning into recently proposed one-bit supervi
sion. Here one-bit supervision denotes a simple Yes or No query about the correc
tness of the model's prediction, and is more efficient than previous active lear
ning methods requiring assigning accurate labels to the queried samples. We clai
m that such one-bit information is intrinsically in accordance with the goal of
contrastive loss that pulls positive pairs together and pushes negative samples
away. Towards this goal, we design an uncertainty metric to actively select samp
les for query. These samples are then fed into different branches according to t
he queried results. The Yes query is treated as positive pairs of the queried ca
tegory for contrastive pulling, while the No query is treated as hard negative p
airs for contrastive repelling. Additionally, we design a negative loss that pen
alizes the negative samples away from the incorrect predicted class, which can b
e treated as optimizing hard negatives for the corresponding category. Our metho
d, termed as ObCP, produces a more powerful active learning framework, and exper
iments on several benchmarks demonstrate its superiority.
****

Estimating Egocentric 3D Human Pose in the Wild With External Weak Supervision
Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, Christian
 Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2022, pp. 13157-13166
Egocentric 3D human pose estimation with a single fisheye camera has drawn a sig
nificant amount of attention recently. However, existing methods struggle with p

ose estimation from in-the-wild images, because they can only be trained on synt
hetic data due to the unavailability of large-scale in-the-wild egocentric datas
ets. Furthermore, these methods easily fail when the body parts are occluded by
or interacting with the surrounding scene. To address the shortage of in-the-wil
d data, we collect a large-scale in-the-wild egocentric dataset called Egocentri
c Poses in the Wild (EgoPW). This dataset is captured by a head-mounted fisheye
camera and an auxiliary external camera, which provides an additional observatio
n of the human body from a third-person perspective during training. We present
a new egocentric pose estimation method, which can be trained on the new dataset
 with weak external supervision. Specifically, we first generate pseudo labels f
or the EgoPW dataset with a spatio-temporal optimization method by incorporating
 the external-view supervision. The pseudo labels are then used to train an egoc
entric pose estimation network. To facilitate the network training, we propose a
 novel learning strategy to supervise the egocentric features with the high-qual
ity features extracted by a pretrained external-view pose estimation model. The
experiments show that our method predicts accurate 3D poses from a single in-the
-wild egocentric image and outperforms the state-of-the-art methods both quantit
atively and qualitatively.

****

Performance-Aware Mutual Knowledge Distillation for Improving Neural Architectur
e Search

Pengtao Xie, Xuefeng Du; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2022, pp. 11922-11932

Knowledge distillation has shown great effectiveness for improving neural archit
ecture search (NAS). Mutual knowledge distillation (MKD), where a group of model
s mutually generate knowledge to train each other, has achieved promising result
s in many applications. In existing MKD methods, mutual knowledge distillation i
s performed between models without scrutiny: a worse-performing model is allowed
 to generate knowledge to train a better-performing model, which may lead to col
lective failures. To address this problem, we propose a performance-aware MKD (P
AMKD) approach for NAS, where knowledge generated by model A is allowed to train
 model B only if the performance of A is better than B. We propose a three-level
 optimization framework to formulate PAMKD, where three learning stages are perf
ormed end-to-end: 1) each model trains an initial model independently; 2) the in
itial models are evaluated on a validation set and better-performing models gene
rate knowledge to train worse-performing models; 3) architectures are updated by
 minimizing a validation loss. Experimental results on a variety of datasets dem
onstrate that our method is effective.

****

Does Text Attract Attention on E-Commerce Images: A Novel Saliency Prediction Da
taset and Method

Lai Jiang, Yifei Li, Shengxi Li, Mai Xu, Se Lei, Yichen Guo, Bo Huang; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 2088-2097

E-commerce images are playing a central role in attracting people's attention wh
en retailing and shopping online, and an accurate attention prediction is of sig
nificant importance for both customers and retailers, where its research is yet
to start. In this paper, we establish the first dataset of saliency e-commerce i
mages (SalECI), which allows for learning to predict saliency on the e-commerce
images. We then provide specialized and thorough analysis by highlighting the di
stinct features of e-commerce images, e.g., non-locality and correlation to text
 regions. Correspondingly, taking advantages of the non-local and self-attention
 mechanisms, we propose a salient SWin-Transformer backbone, followed by a multi
-task learning with saliency and text detection heads, where an information flow
 mechanism is proposed to further benefit both tasks. Experimental results have
verified the state-of-the-art performances of our work in the e-commerce scenari
o.

****

Topologically-Aware Deformation Fields for Single-View 3D Reconstruction

Shivam Duggal, Deepak Pathak; Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2022, pp. 1536-1546
We present a new framework to learn dense 3D reconstruction and correspondence from a single 2D image. The shape is represented implicitly as deformation over a category-level occupancy field and learned in an unsupervised manner from an unaligned image collection without using any 3D supervision. However, image collections usually contain large intra-category topological variation, e.g. images of different chair instances, posing a major challenge. Hence, prior methods are either restricted only to categories with no topological variation for estimating shape and correspondence or focus only on learning shape independently for each instance without any correspondence. To address this issue, we propose a topologically-aware deformation field that maps 3D points in object space to a higher-dimensional canonical space. Given a single image, we first implicitly deform a 3D point in the object space to a learned category-specific canonical space using the topologically-aware field and then learn the 3D shape in the canonical space. Both the canonical shape and deformation field are trained end-to-end using differentiable rendering via learned recurrent ray marcher. Our approach, dubbed TARS, achieves state-of-the-art reconstruction fidelity on several datasets: ShapeNet, Pascal3D+, CUB, and Pix3D chairs.
****
HyperInverter: Improving StyleGAN Inversion via Hypernetwork
Tan M. Dinh, Anh Tuan Tran, Rang Nguyen, Binh-Son Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11389-11398
Real-world image manipulation has achieved fantastic progress in recent years as a result of the exploration and utilization of GAN latent spaces. GAN inversion is the first step in this pipeline, which aims to map the real image to the latent code faithfully. Unfortunately, the majority of existing GAN inversion methods fail to meet at least one of the three requirements listed below: high reconstruction quality, editability, and fast inference. We present a novel two-phase strategy in this research that fits all requirements at the same time. In the first phase, we train an encoder to map the input image to StyleGAN2 W-space, which was proven to have excellent editability but lower reconstruction quality. In the second phase, we supplement the reconstruction ability in the initial phase by leveraging a series of hypernetworks to recover the missing information during inversion. These two steps complement each other to yield high reconstruction quality thanks to the hypernetwork branch and excellent editability due to the inversion done in the W-space. Our method is entirely encoder-based, resulting in extremely fast inference. Extensive experiments on two challenging datasets demonstrate the superiority of our method.
****
Sparse Non-Local CRF
Olga Veksler, Yuri Boykov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4493-4503
CRF is a classical computer vision model which is also useful for deep learning. There are two common CRF types: sparse and dense. Sparse CRF connects only the nearby pixels, while dense CRF has global connectivity. Therefore dense CRF is a more general model, but it is much harder to optimize compared to sparse CRF. In fact, only a certain form of dense CRF is optimized in practice, and even then approximately. We propose a new sparse non-local CRF: it has a sparse number of connections, but it has both local and non-local ones. Like sparse CRF, the total number of connections is small, and our model is easy to optimize exactly. Like dense CRF, our model is more general than sparse CRF due to non-local connections. We show that our sparse non-local CRF can model properties similar to that of the popular Gaussian edge dense CRF. Besides efficiency, another advantage is that our edge weights are less restricted compared to Gaussian edge dense CRF. We design models that take advantage of this flexibility. We also discuss connection of our model to other CRF models. Finally, to prove the usefulness of our model, we evaluate it on the classical application of segmentation from a bounding box and for deep learning based salient object segmentation. We improve state of the art for both applications.

****
Dataset Distillation by Matching Training Trajectories
George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10718-10727
Dataset distillation is the task of synthesizing a small dataset such that a model trained on the synthetic set will match the test accuracy of the model trained on the full dataset. The task is extremely challenging as it often involves backpropagating through the full training process or assuming the strong constraint that a single training step on distilled data can only imitate a single step on real data. In this paper, we propose a new formulation that optimizes our distilled data to guide networks to a similar state as those trained on real data across many training steps. Given a network, we train it for several iterations on our distilled data and optimize the distilled data with respect to the distance between the synthetically trained parameters and the parameters trained on real data. To efficiently obtain the initial and target network parameters for large-scale datasets, we pre-compute and store training trajectories of expert networks trained on the real dataset. Our method handily outperforms existing methods and also allows us to distill with higher-resolution visual data.
****
Towards Driving-Oriented Metric for Lane Detection Models
Takami Sato, Qi Alfred Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17153-17162
After the 2017 TuSimple Lane Detection Challenge, its dataset and evaluation based on accuracy and F1 score have become the de facto standard to measure the performance of lane detection methods. While they have played a major role in improving the performance of lane detection methods, the validity of this evaluation method in downstream tasks has not been adequately researched. In this study, we design 2 new driving-oriented metrics for lane detection: End-to-End Lateral Deviation metric (E2E-LD) is directly formulated based on the requirements of autonomous driving, a core task downstream of lane detection; Per-frame Simulated Lateral Deviation metric (PSLD) is a lightweight surrogate metric of E2E-LD. To evaluate the validity of the metrics, we conduct a large-scale empirical study with 4 major types of lane detection approaches on the TuSimple dataset and our newly constructed dataset Comma2k19-LD. Our results show that the conventional metrics have strongly negative correlations (<=-0.55) with E2E-LD, meaning that some recent improvements purely targeting the conventional metrics may not have led to meaningful improvements in autonomous driving, but rather may actually have made it worse by overfitting to the conventional metrics. On the contrary, PSLD shows statistically significant strong positive correlations (>=0.38) with E2E-LD. As a result, the conventional metrics tend to overestimate less robust models. As autonomous driving is a security/safety-critical system, the underestimation of robustness hinders the sound development of practical lane detection models. We hope that our study will help the community achieve more downstream task-aware evaluations for lane detection.
****
EPro-PnP: Generalized End-to-End Probabilistic Perspective-N-Points for Monocular Object Pose Estimation
Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2781-2790
Locating 3D objects from a single RGB image via Perspective-n-Points (PnP) is a long-standing problem in computer vision. Driven by end-to-end deep learning, recent studies suggest interpreting PnP as a differentiable layer, so that 2D-3D point correspondences can be partly learned by backpropagating the gradient w.r.t. object pose. Yet, learning the entire set of unrestricted 2D-3D points from scratch fails to converge with existing approaches, since the deterministic pose is inherently non-differentiable. In this paper, we propose the EPro-PnP, a probabilistic PnP layer for general end-to-end pose estimation, which outputs a distribution of pose on the SE(3) manifold, essentially bringing categorical Softmax

to the continuous domain. The 2D-3D coordinates and corresponding weights are tr
eated as intermediate variables learned by minimizing the KL divergence between
the predicted and target pose distribution. The underlying principle unifies the
 existing approaches and resembles the attention mechanism. EPro-PnP significant
ly outperforms competitive baselines, closing the gap between PnP-based method a
nd the task-specific leaders on the LineMOD 6DoF pose estimation and nuScenes 3D
 object detection benchmarks.
****

Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection
Yibo Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2022, pp. 7379-7387
In some scenarios, classifier requires detecting out-of-distribution samples far
 from its training data. With desirable characteristics, reconstruction autoenco
der-based methods deal with this problem by using input reconstruction error as
a metric of novelty vs. normality. We formulate the essence of such approach as
a quadruplet domain translation with an intrinsic bias to only query for a proxy
 of conditional data uncertainty. Accordingly, an improvement direction is forma
lized as maximumly compressing the autoencoder's latent space while ensuring its
 reconstructive power for acting as a described domain translator. From it, stra
tegies are introduced including semantic reconstruction, data certainty decompos
ition and normalized L2 distance to substantially improve original methods, whic
h together establish state-of-the-art performance on various benchmarks, e.g., t
he FPR@95%TPR of CIFAR-100 vs. TinyImagenet-crop on Wide-ResNet is 0.2%. Importa
ntly, our method works without any additional data, hard-to-implement structure,
 time-consuming pipeline, and even harming the classification accuracy of known
classes.
****

XYDeblur: Divide and Conquer for Single Image Deblurring
Seo-Won Ji, Jeongmin Lee, Seung-Wook Kim, Jun-Pyo Hong, Seung-Jin Baek, Seung-Wo
n Jung, Sung-Jea Ko; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2022, pp. 17421-17430
Many convolutional neural networks (CNNs) for single image deblurring employ a U
-Net structure to estimate latent sharp images. Having long been proven to be ef
fective in image restoration tasks, a single lane of encoder-decoder architectur
e overlooks the characteristic of deblurring, where a blurry image is generated
from complicated blur kernels caused by tangled motions. Toward an effective net
work architecture, we present complemental sub-solutions learning with a one-enc
oder-two-decoder architecture for single image deblurring. Observing that multip
le decoders successfully learn to decompose information in the encoded features
into directional components, we further improve both the network efficiency and
the deblurring performance by rotating and sharing kernels exploited in the deco
ders, which prevents the decoders from separating unnecessary components such as
 color shift. As a result, our proposed network shows superior results as compar
ed to U-Net while preserving the network parameters, and the use of the proposed
 network as the base network can improve the performance of existing state-of-th
e-art deblurring networks.
****

Generating Diverse and Natural 3D Human Motions From Text
Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, Li Cheng; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2022, pp. 5152-5161
Automated generation of 3D human motions from text is a challenging problem. The
 generated motions are expected to be sufficiently diverse to explore the text-g
rounded motion space, and more importantly, accurately depicting the content in
prescribed text descriptions. Here we tackle this problem with a two-stage appro
ach: text2length sampling and text2motion generation. Text2length involves sampl
ing from the learned distribution function of motion lengths conditioned on the
input text. This is followed by our text2motion module using temporal variationa
l autoencoder to synthesize a diverse set of human motions of the sampled length
s. Instead of directly engaging with pose sequences, we propose motion snippet c

ode as our internal motion representation, which captures local semantic motion contexts and is empirically shown to facilitate the generation of plausible motions faithful to the input text. Moreover, a large-scale dataset of scripted 3D Human motions, HumanML3D, is constructed, consisting of 14,616 motion clips and 44,970 text descriptions. Extensive empirical experiments demonstrate the effectiveness of our approach. Project webpage: https://ericguo5513.github.io/text-to-motion/.

****

E-CIR: Event-Enhanced Continuous Intensity Recovery

Chen Song, Qixing Huang, Chandrajit Bajaj; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7803-7812

A camera begins to sense light the moment we press the shutter button. During the exposure interval, relative motion between the scene and the camera causes motion blur, a common undesirable visual artifact. This paper presents E-CIR, which converts a blurry image into a sharp video represented as a parametric function from time to intensity. E-CIR leverages events as an auxiliary input. We discuss how to exploit the temporal event structure to construct the parametric bases. We demonstrate how to train a deep learning model to predict the function coefficients. To improve the appearance consistency, we further introduce a refinement module to propagate visual features among consecutive frames. Compared to state-of-the-art event-enhanced deblurring approaches, E-CIR generates smoother and more realistic results. The implementation of E-CIR is available at https://github.com/chensong1995/E-CIR.

****

Towards Robust Rain Removal Against Adversarial Attacks: A Comprehensive Benchmark Analysis and Beyond

Yi Yu, Wenhan Yang, Yap-Peng Tan, Alex C. Kot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6013-6022

Rain removal aims to remove rain streaks from images/videos and reduce the disruptive effects caused by rain. It not only enhances image/video visibility but also allows many computer vision algorithms to function properly. This paper makes the first attempt to conduct a comprehensive study on the robustness of deep learning-based rain removal methods against adversarial attacks. Our study shows that, when the image/video is highly degraded, rain removal methods are more vulnerable to the adversarial attacks as small distortions/perturbations become less noticeable or detectable. In this paper, we first present a comprehensive empirical evaluation of various methods at different levels of attacks and with various losses/targets to generate the perturbations from the perspective of human perception and machine analysis tasks. A systematic evaluation of key modules in existing methods is performed in terms of their robustness against adversarial attacks. From the insights of our analysis, we construct a more robust deraining method by integrating these effective modules. Finally, we examine various types of adversarial attacks that are specific to deraining problems and their effects on both human and machine vision tasks, including 1) rain region attacks, adding perturbations only in the rain regions to make the perturbations in the attacked rain images less visible; 2) object-sensitive attacks, adding perturbations only in regions near the given objects. Code is available at https://github.com/yuyi-sd/Robust_Rain_Removal.

****

STCrowd: A Multimodal Dataset for Pedestrian Perception in Crowded Scenes

Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, Yuexin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19608-19617

Accurately detecting and tracking pedestrians in 3D space is challenging due to large variations in rotations, poses and scales. The situation becomes even worse for dense crowds with severe occlusions. However, existing benchmarks either only provide 2D annotations, or have limited 3D annotations with low-density pedestrian distribution, making it difficult to build a reliable pedestrian perception system especially in crowded scenes. To better evaluate pedestrian perception algorithms in crowded scenarios, we introduce a large-scale multimodal dataset,

STCrowd. Specifically, in STCrowd, there are a total of 219K pedestrian instanc es and 20 persons per frame on average, with various levels of occlusion. We pro vide synchronized LiDAR point clouds and camera images as well as their correspo nding 3D labels and joint IDs. STCrowd can be used for various tasks, including LiDAR-only, image-only, and sensor-fusion based pedestrian detection and trackin g. We provide baselines for most of the tasks. In addition, considering the prop erty of sparse global distribution and density-varying local distribution of ped estrians, we further propose a novel method, Density-aware Hierarchical heatmap Aggregation (DHA), to enhance pedestrian perception in crowded scenes. Extensive experiments show that our new method achieves state-of-the-art performance on t he STCrowd dataset, especially on cases with severe occlusion. The dataset and c ode will be released to facilitate related research when the paper is published.

****

Deep Decomposition for Stochastic Normal-Abnormal Transport

Peirong Liu, Yueh Lee, Stephen Aylward, Marc Niethammer; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 187 91-18801

Advection-diffusion equations describe a large family of natural transport proce sses, e.g., fluid flow, heat transfer, and wind transport. They are also used fo r optical flow and perfusion imaging computations. We develop a machine learning model, D^2-SONATA, built upon a stochastic advection-diffusion equation, which predicts the velocity and diffusion fields that drive 2D/3D image time-series of transport. In particular, our proposed model incorporates a model of transport atypicality, which isolates abnormal differences between expected normal transpo rt behavior and the observed transport. In a medical context such a normal-abnor mal decomposition can be used, for example, to quantify pathologies. Specificall y, our model identifies the advection and diffusion contributions from the trans port time-series and simultaneously predicts an anomaly value field to provide a decomposition into normal and abnormal advection and diffusion behavior. To ach ieve improved estimation performance for the velocity and diffusion-tensor field s underlying the advection-diffusion process and for the estimation of the anoma ly fields, we create a 2D/3D anomaly-encoded advection-diffusion simulator, whic h allows for supervised learning. We further apply our model on a brain perfusio n dataset from ischemic stroke patients via transfer learning. Extensive compari sons demonstrate that our model successfully distinguishes stroke lesions (abnor mal) from normal brain regions, while reconstructing the underlying velocity and diffusion tensor fields.

****

Global Context With Discrete Diffusion in Vector Quantised Modelling for Image G eneration

Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, P.N. Suganthan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 11502-11511

The integration of Vector Quantised Variational AutoEncoder (VQ-VAE) with autore gressive models as generation part has yielded high-quality results on image gen eration. However, the autoregressive models will strictly follow the progressive scanning order during the sampling phase. This leads the existing VQ series mod els to hardly escape the trap of lacking global information. Denoising Diffusion Probabilistic Models (DDPM) in the continuous domain have shown a capability to capture the global context, while generating high-quality images. In the discre te state space, some works have demonstrated the potential to perform text gener ation and low resolution image generation. We show that with the help of a conte nt-rich discrete visual codebook from VQ-VAE, the discrete diffusion model can a lso generate high fidelity images with global context, which compensates for the deficiency of the classical autoregressive model along pixel space. Meanwhile, the integration of the discrete VAE with the diffusion model resolves the drawba ck of conventional autoregressive models being oversized, and the diffusion mode l which demands excessive time in the sampling process when generating images. I t is found that the quality of the generated images is heavily dependent on the discrete visual codebook. Extensive experiments demonstrate that the proposed Ve

ctor Quantised Discrete Diffusion Model (VQ-DDM) is able to achieve comparable p
erformance to top-tier methods with low complexity. It also demonstrates outstan
ding advantages over other vectors quantised with autoregressive models in terms
 of image inpainting tasks without additional training.
****

Symmetry and Uncertainty-Aware Object SLAM for 6DoF Object Pose Estimation
Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, X
i Peng, Liu Ren, Guoquan Huang; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2022, pp. 14901-14910
We propose a keypoint-based object-level SLAM framework that can provide globall
y consistent 6DoF pose estimates for symmetric and asymmetric objects alike. To
the best of our knowledge, our system is among the first to utilize the camera p
ose information from SLAM to provide prior knowledge for tracking keypoints on s
ymmetric objects - ensuring that new measurements are consistent with the curren
t 3D scene. Moreover, our semantic keypoint network is trained to predict the Ga
ussian covariance for the keypoints that captures the true error of the predicti
on, and thus is not only useful as a weight for the residuals in the system's op
timization problems, but also as a means to detect harmful statistical outliers
without choosing a manual threshold. Experiments show that our method provides c
ompetitive performance to the state of the art in 6DoF object pose estimation, a
nd at a real-time speed. Our code, pre-trained models, and keypoint labels are a
vailable https://github.com/rpng/suo_slam.
****

AziNorm: Exploiting the Radial Symmetry of Point Cloud for Azimuth-Normalized 3D
 Perception
Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Wenqiang Zhang, Qian Zhang, Chang Hu
ang, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2022, pp. 6387-6396
Studying the inherent symmetry of data is of great importance in machine learnin
g. Point cloud, the most important data format for 3D environmental perception,
is naturally endowed with strong radial symmetry. In this work, we exploit this
radial symmetry via a divide-and-conquer strategy to boost 3D perception perform
ance and ease optimization. We propose Azimuth Normalization (AziNorm), which no
rmalizes the point clouds along the radial direction and eliminates the variabil
ity brought by the difference of azimuth. AziNorm can be flexibly incorporated i
nto most LiDAR-based perception methods. To validate its effectiveness and gener
alization ability, we apply AziNorm in both object detection and semantic segmen
tation. For detection, we integrate AziNorm into two representative detection me
thods, the one-stage SECOND detector and the state-of-the-art two-stage PV-RCNN
detector. Experiments on Waymo Open Dataset demonstrate that AziNorm improves SE
COND and PV-RCNN by 7.03 mAPH and 3.01 mAPH respectively. For segmentation, we i
ntegrate AziNorm into KPConv. On SemanticKitti dataset, AziNorm improves KPConv
by 1.6/1.1 mIoU on val/test set. Besides, AziNorm remarkably improves data effic
iency and accelerates convergence, reducing the requirement of data amounts or t
raining epochs by an order of magnitude. SECOND w/ AziNorm can significantly out
perform fully trained vanilla SECOND, even trained with only 10% data or 10% epo
chs. Code and models are available at https://github.com/hustvl/AziNorm.
****

Towards Multimodal Depth Estimation From Light Fields
Titus Leistner, Radek Mackowiak, Lynton Ardizzone, Ullrich Köthe, Carsten Rother
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 12953-12961
Light field applications, especially light field rendering and depth estimation,
 developed rapidly in recent years. While state-of-the-art light field rendering
 methods handle semi-transparent and reflective objects well, depth estimation m
ethods either ignore these cases altogether or only deliver a weak performance.
We argue that this is due current methods only considering a single "true" depth
, even when multiple objects at different depths contributed to the color of a s
ingle pixel. Based on the simple idea of outputting a posterior depth distributi
on instead of only a single estimate, we develop and explore several different d

eep-learning-based approaches to the problem. Additionally, we contribute the fi rst "multimodal light field depth dataset" that contains the depths of all objec ts which contribute to the color of a pixel. This allows us to supervise the mul timodal depth prediction and also validate all methods by measuring the KL diver gence of the predicted posteriors. With our thorough analysis and novel dataset, we aim to start a new line of depth estimation research that overcomes some of the long-standing limitations of this field.
****

Learning To Recognize Procedural Activities With Distant Supervision
Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, Lore nzo Torresani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 13853-13863
In this paper we consider the problem of classifying fine-grained, multi-step ac tivities (e.g., cooking different recipes, making disparate home improvements, c reating various forms of arts and crafts) from long videos spanning up to severa l minutes. Accurately categorizing these activities requires not only recognizin g the individual steps that compose the task but also capturing their temporal d ependencies. This problem is dramatically different from traditional action clas sification, where models are typically optimized on videos that span only a few seconds and that are manually trimmed to contain simple atomic actions. While st ep annotations could enable the training of models to recognize the individual s teps of procedural activities, existing large-scale datasets in this area do not include such segment labels due to the prohibitive cost of manually annotating temporal boundaries in long videos. To address this issue, we propose to automat ically identify steps in instructional videos by leveraging the distant supervis ion of a textual knowledge base (wikiHow) that includes detailed descriptions of the steps needed for the execution of a wide variety of complex activities. Our method uses a language model to match noisy, automatically-transcribed speech f rom the video to step descriptions in the knowledge base. We demonstrate that vi deo models trained to recognize these automatically-labeled steps (without manua l supervision) yield a representation that achieves superior generalization perf ormance on four downstream tasks: recognition of procedural activities, step cla ssification, step forecasting and egocentric video classification.
****

Multimodal Material Segmentation
Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, Ko Nishino; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1 9800-19808
Recognition of materials from their visual appearance is essential for computer vision tasks, especially those that involve interaction with the real world. Mat erial segmentation, i.e., dense per-pixel recognition of materials, remains chal lenging as, unlike objects, materials do not exhibit clearly discernible visual signatures in their regular RGB appearances. Different materials, however, do le ad to different radiometric behaviors, which can often be captured with non-RGB imaging modalities. We realize multimodal material segmentation from RGB, polari zation, and near-infrared images. We introduce the MCubeS dataset (from MultiMod al Material Segmentation) which contains 500 sets of multimodal images capturing 42 street scenes. Ground truth material segmentation as well as semantic segmen tation are annotated for every image and pixel. We also derive a novel deep neur al network, MCubeSNet, which learns to focus on the most informative combination s of imaging modalities for each material class with a newly derived region-guid ed filter selection (RGFS) layer. We use semantic segmentation, as a prior to "g uide" this filter selection. To the best of our knowledge, our work is the first comprehensive study on truly multimodal material segmentation. We believe our w ork opens new avenues of practical use of material information in safety critica l applications.
****

Multi-Frame Self-Supervised Depth With Transformers
Vitor Guizilini, Rare■ Ambru■, Dian Chen, Sergey Zakharov, Adrien Gaidon; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP

R), 2022, pp. 160-170
Multi-frame depth estimation improves over single-frame approaches by also leveraging geometric relationships between images via feature matching, in addition to learning appearance-based features. In this paper we revisit feature matching for self-supervised monocular depth estimation, and propose a novel transformer architecture for cost volume generation. We use depth-discretized epipolar sampling to select matching candidates, and refine predictions through a series of self- and cross-attention layers. These layers sharpen the matching probability between pixel features, improving over standard similarity metrics prone to ambiguities and local minima. The refined cost volume is decoded into depth estimates, and the whole pipeline is trained end-to-end from videos using only a photometric objective. Experiments on the KITTI and DDAD datasets show that our DepthFormer architecture establishes a new state of the art in self-supervised monocular depth estimation, and is even competitive with highly specialized supervised single-frame architectures. We also show that our learned cross-attention network yields representations transferable across datasets, increasing the effectiveness of pre-training strategies.
****

Weakly Supervised Rotation-Invariant Aerial Object Detection Network
Xiaoxu Feng, Xiwen Yao, Gong Cheng, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14146-14155
Object rotation is among long-standing, yet still unexplored, hard issues encountered in the task of weakly supervised object detection (WSOD) from aerial images. Existing predominant WSOD approaches built on regular CNNs which are not inherently designed to tackle object rotations without corresponding constraints, thereby leading to rotation-sensitive object detector. Meanwhile, current solutions have been prone to fall into the issue with unstable detectors, as they ignore lower-scored instances and may regard them as backgrounds. To address these issues, in this paper, we construct a novel end-to-end weakly supervised Rotation-Invariant aerial object detection Network (RINet). It is implemented with a flexible multi-branch online detector refinement, to be naturally more rotation-perceptive against oriented objects. Specifically, RINet first performs label propagating from the predicted instances to their rotated ones in a progressive refinement manner. Meanwhile, we propose to couple the predicted instance labels among different rotation-perceptive branches for generating rotation-consistent supervision and meanwhile pursuing all possible instances. With the rotation-consistent supervisions, RINet enforces and encourages consistent yet complementary feature learning for WSOD without additional annotations and hyper-parameters. On the challenging NWPU VHR-10.v2 and DIOR datasets, extensive experiments clearly demonstrate that we significantly boost existing WSOD methods to a new state-of-the-art performance. The code will be available at: https://github.com/XiaoxFeng/RINet.
****

Modeling Motion With Multi-Modal Features for Text-Based Video Segmentation
Wangbo Zhao, Kai Wang, Xiangxiang Chu, Fuzhao Xue, Xinchao Wang, Yang You; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11737-11746
Text-based video segmentation aims to segment the target object in a video based on a describing sentence. Incorporating motion information from optical flow maps with appearance and linguistic modalities is crucial yet has been largely ignored by previous work. In this paper, we design a method to fuse and align appearance, motion, and linguistic features to achieve accurate segmentation. Specifically, we propose a multi-modal video transformer, which can fuse and aggregate multi-modal and temporal features between frames. Furthermore, we design a language-guided feature fusion module to progressively fuse appearance and motion features in each feature level with guidance from linguistic features. Finally, a multi-modal alignment loss is proposed to alleviate the semantic gap between features from different modalities. Extensive experiments on A2D Sentences and J-HMDB Sentences verify the performance and the generalization ability of our method compared to the state-of-the-art methods.

****
Surface Reconstruction From Point Clouds by Learning Predictive Context Priors
Baorui Ma, Yu-Shen Liu, Matthias Zwicker, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6326-6337
Surface reconstruction from point clouds is vital for 3D computer vision. State-of-the-art methods leverage large datasets to first learn local context priors that are represented as neural network-based signed distance functions (SDFs) with some parameters encoding the local contexts. To reconstruct a surface at a specific query location at inference time, these methods then match the local reconstruction target by searching for the best match in the local prior space (by optimizing the parameters encoding the local context) at the given query location. However, this requires the local context prior to generalize to a wide variety of unseen target regions, which is hard to achieve. To resolve this issue, we introduce Predictive Context Priors by learning Predictive Queries for each specific point cloud at inference time. Specifically, we first train a local context prior using a large point cloud dataset similar to previous techniques. For surface reconstruction at inference time, however, we specialize the local context prior into our Predictive Context Prior by learning Predictive Queries, which predict adjusted spatial query locations as displacements of the original locations. This leads to a global SDF that fits the specific point cloud the best. Intuitively, the query prediction enables us to flexibly search the learned local context prior over the entire prior space, rather than being restricted to the fixed query locations, and this improves the generalizability. Our method does not require ground truth signed distances, normals, or any additional procedure of signed distance fusion across overlapping regions. Our experimental results in surface reconstruction for single shapes or complex scenes show significant improvements over the state-of-the-art under widely used benchmarks. Code and data are available at https://github.com/mabaorui/PredictableContextPrior.
****
Deformable Video Transformer
Jue Wang, Lorenzo Torresani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14053-14062
Video transformers have recently emerged as an effective alternative to convolutional networks for action classification. However, most prior video transformers adopt either global space-time attention or hand-defined strategies to compare patches within and across frames. These fixed attention schemes not only have high computational cost but, by comparing patches at predetermined locations, they neglect the motion dynamics in the video. In this paper, we introduce the Deformable Video Transformer (DVT), which dynamically predicts a small subset of video patches to attend for each query location based on motion information, thus allowing the model to decide where to look in the video based on correspondences across frames. Crucially, these motion-based correspondences are obtained at zero-cost from information stored in the compressed format of the video. Our deformable attention mechanism is optimized directly with respect to classification performance, thus eliminating the need for suboptimal hand-design of attention strategies. Experiments on four large-scale video benchmarks (Kinetics-400, Something-Something-V2, EPIC-KITCHENS and Diving-48) demonstrate that, compared to existing video transformers, our model achieves higher accuracy at the same or lower computational cost, and it attains state-of-the-art results on these four datasets.
****
Self-Supervised Keypoint Discovery in Behavioral Videos
Jennifer J. Sun, Serim Ryou, Roni H. Goldshmid, Brandon Weissbourd, John O. Dabiri, David J. Anderson, Ann Kennedy, Yisong Yue, Pietro Perona; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2171-2180
We propose a method for learning the posture and structure of agents from unlabelled behavioral videos. Starting from the observation that behaving agents are generally the main sources of movement in behavioral videos, our method, Behavior

al Keypoint Discovery (B-KinD), uses an encoder-decoder architecture with a geometric bottleneck to reconstruct the spatiotemporal difference between video frames. By focusing only on regions of movement, our approach works directly on input videos without requiring manual annotations. Experiments on a variety of agent types (mouse, fly, human, jellyfish, and trees) demonstrate the generality of our approach and reveal that our discovered keypoints represent semantically meaningful body parts, which achieve state-of-the-art performance on keypoint regression among self-supervised methods. Additionally, B-KinD achieve comparable performance to supervised keypoints on downstream tasks, such as behavior classification, suggesting that our method can dramatically reduce model training costs vis-a-vis supervised methods.

****

IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes

Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2822-2831

Indoor scenes exhibit significant appearance variations due to myriad interactions between arbitrarily diverse object shapes, spatially-changing materials, and complex lighting. Shadows, highlights, and inter-reflections caused by visible and invisible light sources require reasoning about long-range interactions for inverse rendering, which seeks to recover the components of image formation, namely, shape, material, and lighting. In this work, our intuition is that the long-range attention learned by transformer architectures is ideally suited to solve longstanding challenges in single-image inverse rendering. We demonstrate with a specific instantiation of a dense vision transformer, \Ours , that excels at both single-task and multi-task reasoning required for inverse rendering. Specifically, we propose a transformer architecture to simultaneously estimate depths, normals, spatially-varying albedo, roughness and lighting from a single image of an indoor scene. Our extensive evaluations on benchmark datasets demonstrate state-of-the-art results on each of the above tasks, enabling applications like object insertion and material editing in a single unconstrained real image, with greater photorealism than prior works. Code and data are publicly released at https://github.com/ViLab-UCSD/IRISformer

****

DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation

Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Ça■lar ■enaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, Laura Leal-Taixé; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21158-21167

Earth observation is a fundamental tool for monitoring the evolution of land use in specific areas of interest. Observing and precisely defining change, in this context, requires both time-series data and pixel-wise segmentations. To that end, we propose the DynamicEarthNet dataset that consists of daily, multi-spectral satellite observations of 75 selected areas of interest distributed over the globe with imagery from Planet Labs. These observations are paired with pixel-wise monthly semantic segmentation labels of 7 land use and land cover (LULC) classes. DynamicEarthNet is the first dataset that provides this unique combination of daily measurements and high-quality labels. In our experiments, we compare several established baselines that either utilize the daily observations as additional training data (semi-supervised learning) or multiple observations at once (spatio-temporal learning) as a point of reference for future research. Finally, we propose a new evaluation metric SCS that addresses the specific challenges associated with time-series semantic change segmentation. The data is available at: https://mediatum.ub.tum.de/1650201.

****

Connecting the Complementary-View Videos: Joint Camera Identification and Subject Association

Ruize Han, Yiyang Gan, Jiacheng Li, Feifan Wang, Wei Feng, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2416-2425
We attempt to connect the data from complementary views, i.e., top view from drone-mounted cameras in the air, and side view from wearable cameras on the ground. Collaborative analysis of such complementary-view data can facilitate to build the air-ground cooperative visual system for various kinds of applications. This is a very challenging problem due to the large view difference between top and side views. In this paper, we develop a new approach that can simultaneously handle three tasks: i) localizing the side-view camera in the top view; ii) estimating the view direction of the side-view camera; iii) detecting and associating the same subjects on the ground across the complementary views. Our main idea is to explore the spatial position layout of the subjects in two views. In particular, we propose a spatial-aware position representation method to embed the spatial-position distribution of the subjects in different views. We further design a cross-view video collaboration framework composed of a camera identification module and a subject association module to simultaneously perform the above three tasks. We collect a new synthetic dataset consisting of top-view and side-view video sequence pairs for performance evaluation and the experimental results show the effectiveness of the proposed method.
****

End-to-End Trajectory Distribution Prediction Based on Occupancy Grid Maps
Ke Guo, Wenxi Liu, Jia Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2242-2251
In this paper, we aim to forecast a future trajectory distribution of a moving agent in the real world, given the social scene images and historical trajectories. Yet, it is a challenging task because the ground-truth distribution is unknown and unobservable, while only one of its samples can be applied for supervising model learning, which is prone to bias. Most recent works focus on predicting diverse trajectories in order to cover all modes of the real distribution, but they may despise the precision and thus give too much credit to unrealistic predictions. To address the issue, we learn the distribution with symmetric cross-entropy using occupancy grid maps as an explicit and scene-compliant approximation to the ground-truth distribution, which can effectively penalize unlikely predictions. In specific, we present an inverse reinforcement learning based multi-modal trajectory distribution forecasting framework that learns to plan by an approximate value iteration network in an end-to-end manner. Besides, based on the predicted distribution, we generate a small set of representative trajectories through a differentiable Transformer-based network, whose attention mechanism helps to model the relations of trajectories. In experiments, our method achieves state-of-the-art performance on the Stanford Drone Dataset and Intersection Drone Dataset.
****

Fast, Accurate and Memory-Efficient Partial Permutation Synchronization
Shaohan Li, Yunpeng Shi, Gilad Lerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15735-15743
Previous partial permutation synchronization (PPS) algorithms, which are commonly used for multi-object matching, often involve computation-intensive and memory-demanding matrix operations. These operations become intractable for large scale structure-from-motion datasets. For pure permutation synchronization, the recent Cycle-Edge Message Passing (CEMP) framework suggests a memory-efficient and fast solution. Here we overcome the restriction of CEMP to compact groups and propose an improved algorithm, CEMP-Partial, for estimating the corruption levels of the observed partial permutations. It allows us to subsequently implement a nonconvex weighted projected power method without the need of spectral initialization. The resulting new PPS algorithm, MatchFAME (Fast, Accurate and Memory-Efficient Matching), only involves sparse matrix operations, and thus enjoys lower time and space complexities in comparison to previous PPS algorithms. We prove that under adversarial corruption, though without additive noise and with certain assumptions, CEMP-Partial is able to exactly classify corrupted and clean partial

permutations. We demonstrate the state-of-the-art accuracy, speed and memory efficiency of our method on both synthetic and real datasets.
****

Quantization-Aware Deep Optics for Diffractive Snapshot Hyperspectral Imaging
Lingen Li, Lizhi Wang, Weitao Song, Lei Zhang, Zhiwei Xiong, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19780-19789
Diffractive snapshot hyperspectral imaging based on the deep optics framework has been striving to capture the spectral images of dynamic scenes. However, existing deep optics frameworks all suffer from the mismatch between the optical hardware and the reconstruction algorithm due to the quantization operation in the diffractive optical element (DOE) fabrication, leading to the limited performance of hyperspectral imaging in practice. In this paper, we propose the quantization-aware deep optics for diffractive snapshot hyperspectral imaging. Our key observation is that common lithography techniques used in fabricating DOEs need to quantize the DOE height map to a few levels, and can freely set the height for each level. Therefore, we propose to integrate the quantization operation into the DOE height map optimization and design an adaptive mechanism to adjust the physical height of each quantization level. According to the optimization, we fabricate the quantized DOE directly and build a diffractive hyperspectral snapshot imaging system. Our method develops the deep optics framework to be more practical through the awareness of and adaptation to the quantization operation of the DOE physical structure, making the fabricated DOE and the reconstruction algorithm match each other systematically. Extensive synthetic simulation and real hardware experiments validate the superior performance of our method.
****

Weakly Supervised Temporal Action Localization via Representative Snippet Knowledge Propagation
Linjiang Huang, Liang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3272-3281
Weakly supervised temporal action localization targets at localizing temporal boundaries of actions and simultaneously identify their categories with only video-level category labels. Many existing methods seek to generate pseudo labels for bridging the discrepancy between classification and localization, but usually only make use of limited contextual information for pseudo label generation. To alleviate this problem, we propose a representative snippet summarization and propagation framework. Our method seeks to mine the representative snippets in each video for better propagating information between video snippets. For each video, its own representative snippets and the representative snippets from a memory bank are propagated to update the input features in an intra- and inter-video manner. The pseudo labels are generated from the temporal class activation maps of the updated features to rectify the predictions of the main branch. Our method obtains superior performance in comparison to the existing methods on two benchmarks, THUMOS14 and ActivityNet1.3, achieving gains as high as 1.2% in terms of average mAP on THUMOS14. Our code is available at https://github.com/LeonHLJ/RSKP.
****

Parametric Scattering Networks
Shanel Gauthier, Benjamin Thérien, Laurent Alsène-Racicot, Muawiz Chaudhary, Irina Rish, Eugene Belilovsky, Michael Eickenberg, Guy Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5749-5758
The wavelet scattering transform creates geometric invariants and deformation stability. In multiple signal domains, it has been shown to yield more discriminative representations compared to other non-learned representations and to outperform learned representations in certain tasks, particularly on limited labeled data and highly structured signals. The wavelet filters used in the scattering transform are typically selected to create a tight frame via a parameterized mother wavelet. In this work, we investigate whether this standard wavelet filterbank construction is optimal. Focusing on Morlet wavelets, we propose to learn the sc

ales, orientations, and aspect ratios of the filters to produce problem-specific parameterizations of the scattering transform. We show that our learned versions of the scattering transform yield significant performance gains in small-sample classification settings over the standard scattering transform. Moreover, our empirical results suggest that traditional filterbank constructions may not always be necessary for scattering transforms to extract effective representations.
****

SketchEdit: Mask-Free Local Image Manipulation With Partial Sketches
Yu Zeng, Zhe Lin, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5951-5961
Sketch-based image manipulation is an interactive image editing task to modify an image based on input sketches from users. Existing methods typically convert this task into a conditional inpainting problem, which requires an additional mask from users indicating the region to modify. Then the masked regions are regarded as missing and filled by an inpainting model conditioned on the sketch. With this formulation, paired training data can be easily obtained by randomly creating masks and extracting edges or contours. Although this setup simplifies data preparation and model design, it complicates user interaction and discards useful information in masked regions. To this end, we propose a new framework for sketch-based image manipulation that only requires sketch inputs from users and utilizes the entire original image. Given an image and sketch, our model automatically predicts the target modification region and encodes it into a structure agnostic style vector. A generator then synthesizes the new image content based on the style vector and sketch. The manipulated image is finally produced by blending the generator output into the modification region of the original image. Our model can be trained in a self-supervised fashion by learning the reconstruction of an image region from the style vector and sketch. The proposed framework offers simpler and more intuitive user workflows for sketch-based image manipulation and provides better results than previous approaches. The code and interactive demo can be found in the supplementary material.
****

ScaleNet: A Shallow Architecture for Scale Estimation
Axel Barroso-Laguna, Yurun Tian, Krystian Mikolajczyk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12808-12818
In this paper, we address the problem of estimating scale factors between images. We formulate the scale estimation problem as a prediction of a probability distribution over scale factors. We design a new architecture, ScaleNet, that exploits dilated convolutions as well as self- and cross-correlation layers to predict the scale between images. We demonstrate that rectifying images with estimated scales leads to significant performance improvements for various tasks and methods. Specifically, we show how ScaleNet can be combined with sparse local features and dense correspondence networks to improve camera pose estimation, 3D reconstruction, or dense geometric matching in different benchmarks and datasets. We provide an extensive evaluation on several tasks, and analyze the computational overhead of ScaleNet. The code, evaluation protocols, and trained models are publicly available.
****

E2EC: An End-to-End Contour-Based Method for High-Quality High-Speed Instance Segmentation
Tao Zhang, Shiqing Wei, Shunping Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4443-4452
Contour-based instance segmentation methods have developed rapidly recently but feature rough and handcrafted front-end contour initialization, which restricts the model performance, and an empirical and fixed backend predicted-label vertex pairing, which contributes to the learning difficulty. In this paper, we introduce a novel contour-based method, named E2EC, for high-quality instance segmentation. Firstly, E2EC applies a novel learnable contour initialization architecture instead of handcrafted contour initialization. This consists of a contour initialization module for constructing more explicit learning goals and a global con

tour deformation module for taking advantage of all of the vertices' features be tter. Secondly, we propose a novel label sampling scheme, named multi-direction alignment, to reduce the learning difficulty. Thirdly, to improve the quality of the boundary details, we dynamically match the most appropriate predicted-groun d truth vertex pairs and propose the corresponding loss function named dynamic m atching loss. The experiments showed that E2EC can achieve a state-of-the-art pe rformance on the KITTI INStance (KINS) dataset, the Semantic Boundaries Dataset (SBD), the Cityscapes and the COCO dataset. E2EC is also efficient for use in re al-time applications, with an inference speed of 36 fps for 512x512 images on an NVIDIA A6000 GPU. Code will be released at https://github.com/zhang-tao-whu/e2e c.

****

Bounded Adversarial Attack on Deep Content Features

Qiuling Xu, Guanhong Tao, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15203-15212

We propose a novel adversarial attack targeting content features in some deep la yer, that is, individual neurons in the layer. A naive method that enforces a fi xed value/percentage bound for neuron activation values can hardly work and gene rates very noisy samples. The reason is that the level of perceptual variation e ntailed by a fixed value bound is non-uniform across neurons and even for the sa me neuron. We hence propose a novel distribution quantile bound for activation v alues and a polynomial barrier loss function. Given a benign input, a fixed quan tile bound is translated to many value bounds, one for each neuron, based on the distributions of the neuron's activations and the current activation value on t he given input. These individualized bounds enable fine-grained regulation, allo wing content feature mutations with bounded perceptional variations. Our evaluat ion on ImageNet and five different model architectures demonstrates that our att ack is effective. Compared to seven other latest adversarial attacks in both the pixel space and the feature space, our attack can achieve the state-of-the-art trade-off between attack success rate and imperceptibility.

****

BatchFormer: Learning To Explore Sample Relationships for Robust Representation Learning

Zhi Hou, Baosheng Yu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2022, pp. 7256-7266

Despite the success of deep neural networks, there are still many challenges in deep representation learning due to the data scarcity issues such as data imbala nce, unseen distribution, and domain shift. To address the above-mentioned issue s, a variety of methods have been devised to explore the sample relationships in a vanilla way (i.e., from the perspectives of either the input or the loss func tion), failing to explore the internal structure of deep neural networks for lea rning with sample relationships. Inspired by this, we propose to enable deep neu ral networks themselves with the ability to learn the sample relationships from each mini-batch. Specifically, we introduce a batch transformer module or BatchF ormer, which is then applied into the batch dimension of each mini-batch to impl icitly explore sample relationships during training. By doing this, the proposed method enables the collaboration of different samples, e.g., the head-class sam ples can also contribute to the learning of the tail classes for long-tailed rec ognition. Furthermore, to mitigate the gap between training and testing, we shar e the classifier between with or without the BatchFormer during training, which can thus be removed during testing. We perform extensive experiments on over ten datasets and the proposed method achieves significant improvements on different data scarcity applications without any bells and whistles, including the tasks of long-tailed recognition, compositional zero-shot learning, domain generalizat ion, and contrastive learning. Code is made publicly available at https://github .com/zhihou7/BatchFormer.

****

Self-Supervised Image-Specific Prototype Exploration for Weakly Supervised Seman tic Segmentation

Qi Chen, Lingxiao Yang, Jian-Huang Lai, Xiaohua Xie; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4288-4298

Weakly Supervised Semantic Segmentation (WSSS) based on image-level labels has attracted much attention due to low annotation costs. Existing methods often rely on Class Activation Mapping (CAM) that measures the correlation between image pixels and classifier weight. However, the classifier focuses only on the discriminative regions while ignoring other useful information in each image, resulting in incomplete localization maps. To address this issue, we propose a Self-supervised Image-specific Prototype Exploration (SIPE) that consists of an Image-specific Prototype Exploration (IPE) and a General-Specific Consistency (GSC) loss. Specifically, IPE tailors prototypes for every image to capture complete regions, formed our Image-Specific CAM (IS-CAM). GSC is proposed to construct the consistency of general CAM and our specific IS-CAM, which further optimizes the feature representation and empowers a self-correction ability of prototype exploration. Extensive experiments are conducted on PASCAL VOC 2012 and MS COCO 2014 segmentation benchmark and results show our SIPE achieves new state-of-the-art performance using only image-level labels. The code is available at https://github.com/chenqi1126/SIPE.

****

CAD: Co-Adapting Discriminative Features for Improved Few-Shot Classification

Philip Chikontwe, Soopil Kim, Sang Hyun Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14554-14563

Few-shot classification is a challenging problem that aims to learn a model that can adapt to unseen classes given a few labeled samples. Recent approaches pre-train a feature extractor, and then fine-tune for episodic meta-learning. Other methods leverage spatial features to learn pixel-level correspondence while jointly training a classifier. However, results using such approaches show marginal improvements. In this paper, inspired by the transformer style self-attention mechanism, we propose a strategy to cross-attend and re-weight discriminative features for few-shot classification. Given a base representation of support and query images after global pooling, we introduce a single shared module that projects features and cross-attends in two aspects: (i) query to support, and (ii) support to query. The module computes attention scores between features to produce an attention pooled representation of features in the same class that is later added to the original representation followed by a projection head. This effectively re-weights features in both aspects (i & ii) to produce features that better facilitate improved metric-based meta-learning. Extensive experiments on public benchmarks show our approach outperforms state-of-the-art methods by 3% 5%.

****

Fingerprinting Deep Neural Networks Globally via Universal Adversarial Perturbations

Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, Minhui Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13430-13439

In this paper, we propose a novel and practical mechanism which enables the service provider to verify whether a suspect model is stolen from the victim model via model extraction attacks. Our key insight is that the profile of a DNN model's decision boundary can be uniquely characterized by its Universal Adversarial Perturbations (UAPs). UAPs belong to a low-dimensional subspace and piracy models' subspaces are more consistent with victim model's subspace compared with non-piracy model. Based on this, we propose a UAP fingerprinting method for DNN models and train an encoder via contrastive learning that takes fingerprint as inputs, outputs a similarity score. Extensive studies show that our framework can detect model IP breaches with confidence > 99.99% within only 20 fingerprints of the suspect model. It has good generalizability across different model architectures and is robust against post-modifications on stolen models.

****

Learning Multi-View Aggregation in the Wild for Large-Scale 3D Semantic Segmentation

Damien Robert, Bruno Vallet, Loic Landrieu; Proceedings of the IEEE/CVF Conferen

ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5575-5584
Recent works on 3D semantic segmentation propose to exploit the synergy between images and point clouds by processing each modality with a dedicated network and projecting learned 2D features onto 3D points. Merging large-scale point clouds and images raises several challenges, such as constructing a mapping between points and pixels, and aggregating features between multiple views. Current methods require mesh reconstruction or specialized sensors to recover occlusions, and use heuristics to select and aggregate available images. In contrast, we propose an end-to-end trainable multi-view aggregation model leveraging the viewing conditions of 3D points to merge features from images taken at arbitrary positions. Our method can combine standard 2D and 3D networks and outperforms both 3D models operating on colorized point clouds and hybrid 2D/3D networks without requiring colorization, meshing, or true depth maps. We set a new state-of-the-art for large-scale indoor/outdoor semantic segmentation on S3DIS (74.7 mIoU 6-Fold) and on KITTI-360 (58.3 mIoU). Our full pipeline is accessible at https://github.com/drprojects/DeepViewAgg, and only requires raw 3D scans and a set of images and poses.
****

ManiTrans: Entity-Level Text-Guided Image Manipulation via Token-Wise Semantic Alignment and Generation
Jianan Wang, Guansong Lu, Hang Xu, Zhenguo Li, Chunjing Xu, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10707-10717
Existing text-guided image manipulation methods aim to modify the appearance of the image or to edit a few objects in a virtual or simple scenario, which is far from practical application. In this work, we study a novel task on text-guided image manipulation on the entity level in the real world. The task imposes three basic requirements, (1) to edit the entity consistent with the text descriptions, (2) to preserve the text-irrelevant regions, and (3) to merge the manipulated entity into the image naturally. To this end, we propose a new transformer-based framework based on the two-stage image synthesis method, namely ManiTrans, which can not only edit the appearance of entities but also generate new entities corresponding to the text guidance. Our framework incorporates a semantic alignment module to locate the image regions to be manipulated, and a semantic loss to help align the relationship between the vision and language. We conduct extensive experiments on the real datasets, CUB, Oxford, and COCO datasets to verify that our method can distinguish the relevant and irrelevant regions and achieve more precise and flexible manipulation compared with baseline methods.
****

Improving Video Model Transfer With Dynamic Representation Learning
Yi Li, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19280-19291
Temporal modeling is an essential element in video understanding. While deep convolution-based architectures have been successful at solving large-scale video recognition datasets, recent work has pointed out that they are biased towards modeling short-range relations, often failing to capture long-term temporal structures in the videos, leading to poor transfer and generalization to new datasets. In this work, the problem of dynamic representation learning (DRL) is studied. We propose dynamic score, a measure of video dynamic modeling that describes the additional amount of information learned by a video network that cannot be captured by pure spatial student through knowledge distillation. DRL is then formulated as an adversarial learning problem between the video and spatial models, with the objective of maximizing the dynamic score of learned spatiotemporal classifier. The quality of learned video representations are evaluated on a diverse set of transfer learning problems concerning many-shot and few-shot action classification. Experimental results show that models learned with DRL outperform baselines in dynamic modeling, demonstrating higher transferability and generalization capacity to novel domains and tasks.
****

PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposi

tion
Partha Das, Sezer Karaoglu, Theo Gevers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19790-19799
Intrinsic image decomposition is the process of recovering the image formation components (reflectance and shading) from an image. Previous methods employ either explicit priors to constrain the problem or implicit constraints as formulated by their losses (deep learning). These methods can be negatively influenced by strong illumination conditions causing shading-reflectance leakages. Therefore, in this paper, an end-to-end edge-driven hybrid CNN approach is proposed for intrinsic image decomposition. Edges correspond to illumination invariant gradients. To handle hard negative illumination transitions, a hierarchical approach is taken including global and local refinement layers. We make use of attention layers to further strengthen the learning process. An extensive ablation study and large scale experiments are conducted showing that it is beneficial for edge-driven hybrid IID networks to make use of illumination invariant descriptors and that separating global and local cues helps in improving the performance of the network. Finally, it is shown that the proposed method obtains state of the art performance and is able to generalise well to real world images. The project page with pretrained models, finetuned models and network code can be found at: https://ivi.fnwi.uva.nl/cv/pienet/
****

Clothes-Changing Person Re-Identification With RGB Modality Only
Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1060-1069
The key to address clothes-changing person re-identification (re-id) is to extract clothes-irrelevant features, e.g., face, hairstyle, body shape, and gait. Most current works mainly focus on modeling body shape from multi-modality information (e.g., silhouettes and sketches), but do not make full use of the clothes-irrelevant information in the original RGB images. In this paper, we propose a Clothes-based Adversarial Loss (CAL) to mine clothes-irrelevant features from the original RGB images by penalizing the predictive power of re-id model w.r.t. clothes. Extensive experiments demonstrate that using RGB images only, CAL outperforms all state-of-the-art methods on widely-used clothes-changing person re-id benchmarks. Besides, compared with images, videos contain richer appearance and additional temporal information, which can be used to model proper spatiotemporal patterns to assist clothes-changing re-id. Since there is no publicly available clothes-changing video re-id dataset, we contribute a new dataset named CCVID and show that there exists much room for improvement in modeling spatiotemporal information. The code and new dataset are available at: https://github.com/guxinqian/Simple-CCReID.
****

Chitransformer: Towards Reliable Stereo From Cues
Qing Su, Shihao Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1939-1949
Current stereo matching techniques are challenged by restricted searching space, occluded regions and sheer size. While single image depth estimation is spared from these challenges and can achieve satisfactory results with the extracted monocular cues, the lack of stereoscopic relationship renders the monocular prediction less reliable on its own, especially in highly dynamic or cluttered environments. To address these issues in both scenarios, we present an optic-chiasm-inspired self-supervised binocular depth estimation method, wherein a vision transformer (ViT) with gated positional cross-attention (GPCA) layers is designed to enable feature-sensitive pattern retrieval between views while retaining the extensive context information aggregated through self-attentions. Monocular cues from a single view are thereafter conditionally rectified by a blending layer with the retrieved pattern pairs. This crossover design is biologically analogous to the optic-chasma structure in the human visual system and hence the name, ChiTransformer. Our experiments show that this architecture yields substantial improvements over state-of-the-art self-supervised stereo approaches by 11%, and can be

used on both rectilinear and non-rectilinear (e.g., fisheye) images.
****

Robust Image Forgery Detection Over Online Social Network Shared Images
Haiwei Wu, Jiantao Zhou, Jinyu Tian, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13440-13449
The increasing abuse of image editing softwares, such as Photoshop and Meitu, causes the authenticity of digital images questionable. Meanwhile, the widespread availability of online social networks (OSNs) makes them the dominant channels for transmitting forged images to report fake news, propagate rumors, etc. Unfortunately, various lossy operations adopted by OSNs, e.g., compression and resizing, impose great challenges for implementing the robust image forgery detection. To fight against the OSN-shared forgeries, in this work, a novel robust training scheme is proposed. We first conduct a thorough analysis of the noise introduced by OSNs, and decouple it into two parts, i.e., predictable noise and unseen noise, which are modelled separately. The former simulates the noise introduced by the disclosed (known) operations of OSNs, while the latter is designed to not only complete the previous one, but also take into account the defects of the detector itself. We then incorporate the modelled noise into a robust training framework, significantly improving the robustness of the image forgery detector. Extensive experimental results are presented to validate the superiority of the proposed scheme compared with several state-of-the-art competitors. Finally, to promote the future development of the image forgery detection, we build a public forgeries dataset based on four existing datasets and three most popular OSNs. The designed detector recently won the top ranking in a certificate forgery detection competition. The source code and dataset are available at https://github.com/HighwayWu/ImageForensicsOSN.
****

QS-Attn: Query-Selected Attention for Contrastive Learning in I2I Translation
Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, Qingli Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18291-18300
Unpaired image-to-image (I2I) translation often requires to maximize the mutual information between the source and the translated images across different domains, which is critical for the generator to keep the source content and prevent it from unnecessary modifications. The self-supervised contrastive learning has already been successfully applied in the I2I. By constraining features from the same location to be closer than those from different ones, it implicitly ensures the result to take content from the source. However, previous work uses the features from random locations to impose the constraint, which may not be appropriate since some locations contain less information of source domain. Moreover, the feature itself does not reflect the relation with others. This paper deals with these problems by intentionally selecting significant anchor points for contrastive learning. We design a query-selected attention (QS-Attn) module, which compares feature distances in the source domain, giving an attention matrix with a probability distribution in each row. Then we select queries according to their measurement of significance, computed from the distribution. The selected ones are regarded as anchors for contrastive loss. At the same time, the reduced attention matrix is employed to route features in both domains, so that source relations maintain in the synthesis. We validate our proposed method in three different I2I datasets, showing that it increases the image quality without adding learnable parameters.
****

Physically Disentangled Intra- and Inter-Domain Adaptation for Varicolored Haze Removal
Yi Li, Yi Chang, Yan Gao, Changfeng Yu, Luxin Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5841-5850
Learning-based image dehazing methods have achieved marvelous progress during the past few years. On one hand, most approaches heavily rely on synthetic data and may face difficulties to generalize well in real scenes, due to the huge domain gap between synthetic and real images. On the other hand, very few works have

considered the varicolored haze, caused by chromatic casts in real scenes. In th is work, our goal is to handle the new task: real-world varicolored haze removal . To this end, we propose a physically disentangled joint intra- and inter-domai n adaptation paradigm, in which intra-domain adaptation focuses on color correct ion and inter-domain procedure transfers knowledge between synthetic and real do mains. We first learn to physically disentangle haze images into three component s complying with the scattering model: background, transmission map, and atmosph eric light. Since haze color is determined by atmospheric light, we perform intr a-domain adaptation by specifically translating atmospheric light from varicolor ed space to unified color-balanced space, and then reconstructing color-balanced haze image through the scattering model. Consequently, we perform inter-domain adaptation between the synthetic and real images by mutually exchanging the back ground and other two components. Then we can reconstruct both identity and domai n-translated haze images with self-consistency and adversarial loss. Extensive e xperiments demonstrate the superiority of the proposed method over the state-of- the-art for real varicolored image dehazing.

****

Modality-Agnostic Learning for Radar-Lidar Fusion in Vehicle Detection

Yu-Jhe Li, Jinhyung Park, Matthew O'Toole, Kris Kitani; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 918-927

Fusion of multiple sensor modalities such as camera, Lidar, and Radar, which are commonly found on autonomous vehicles, not only allows for accurate detection b ut also robustifies perception against adverse weather conditions and individual sensor failures. Due to inherent sensor characteristics, Radar performs well un der extreme weather conditions (snow, rain, fog) that significantly degrade came ra and Lidar. Recently, a few works have developed vehicle detection methods fus ing Lidar and Radar signals, i.e., MVDNet. However, these models are typically d eveloped under the assumption that the models always have access to two error-fr ee sensor streams. If one of the sensors is unavailable or missing, the model ma y fail catastrophically. To mitigate this problem, we propose the Self-Training Multimodal Vehicle Detection Network (ST-MVDNet) which leverages a Teacher-Stude nt mutual learning framework and a simulated sensor noise model used in strong d ata augmentation for Lidar and Radar. We show that by (1) enforcing output consi stency between a Teacher network and a Student network and by (2) introducing mi ssing modalities (strong augmentations) during training, our learned model break s away from the error-free sensor assumption. This consistency enforcement enabl es the Student model to handle missing data properly and improve the Teacher mod el by updating it with the Student model's exponential moving average. Our exper iments demonstrate that our proposed learning framework for multi-modal detectio n is able to better handle missing sensor data during inference. Furthermore, ou r method achieves new state-of-the-art performance (5% gain) on the Oxford Radar Robotcar dataset under various evaluation settings.

****

A Re-Balancing Strategy for Class-Imbalanced Classification Based on Instance Di fficulty

Sihao Yu, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Zizhen Wang, Xueqi Cheng; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 70-79

Real-world data often exhibits class-imbalanced distributions, where a few class es (a.k.a. majority classes) occupy most instances and lots of classes (a.k.a. m inority classes) have few instances. Neural classification models usually perfor m poorly on minority classes when training on such imbalanced datasets. To impro ve the performance on minority classes, existing methods typically re-balance th e data distribution at the class level, i.e., assigning higher weights to minori ty classes and lower weights to majority classes during the training process. Ho wever, we observe that even the majority classes contain difficult instances to learn. By reducing the weights of the majority classes, such instances would bec ome more difficult to learn and hurt the overall performance consequently. To ta ckle this problem, we propose a novel instance-level re-balancing strategy, whic

h dynamically adjusts the sampling probabilities of instances according to the i
nstance difficulty. Here the instance difficulty is measured based on the learni
ng speed of instance, which is inspired by the human-leaning process (i.e., easi
er instances will be learned faster). We theoretically prove the correctness and
 convergence of our re-sampling algorithm. Empirical experiments demonstrate tha
t our method significantly outperforms state-of-the-art re-balancing methods on
the class-imbalanced datasets.
****

Representation Compensation Networks for Continual Semantic Segmentation
Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, Ming-Ming Cheng; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 7053-7064
In this work, we study the continual semantic segmentation problem, where the de
ep neural networks are required to incorporate new classes continually without c
atastrophic forgetting. We propose to use a structural re-parameterization mecha
nism, named representation compensation (RC) module, to decouple the representat
ion learning of both old and new knowledge. The RC module consists of two dynami
cally evolved branches with one frozen and one trainable. Besides, we design a p
ooled cube knowledge distillation strategy on both spatial and channel dimension
s to further enhance the plasticity and stability of the model. We conduct exper
iments on two challenging continual semantic segmentation scenarios, continual c
lass segmentation and continual domain segmentation. Without any extra computati
onal overhead and parameters during inference, our method outperforms state-of-t
he-art performance. The code is available at https://github.com/zhangchbin/RCIL.
****

Adaptive Gating for Single-Photon 3D Imaging
Ryan Po, Adithya Pediredla, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16354-16363
Single-photon avalanche diodes (SPADs) are growing in popularity for depth sensi
ng tasks. However, SPADs still struggle in the presence of high ambient light du
e to the effects of pile-up. Conventional techniques leverage fixed or asynchron
ous gating to minimize pile-up effects, but these gating schemes are all non-ada
ptive, as they are unable to incorporate factors such as scene priors and previo
us photon detections into their gating strategy. We propose an adaptive gating s
cheme built upon Thompson sampling. Adaptive gating periodically updates the gat
e position based on prior photon observations in order to minimize depth errors.
 Our experiments show that our gating strategy results in significantly reduced
depth reconstruction error and acquisition time, even when operating outdoors un
der strong sunlight conditions.
****

Tracking People by Predicting 3D Appearance, Location and Pose
Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Jitendra Malik; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2022, pp. 2740-2749
We present an approach for tracking people in monocular videos by predicting the
ir future 3D representations. To achieve this, we first lift people to 3D from a
 single frame in a robust manner. This lifting includes information about the 3D
 pose of the person, their location in the 3D space, and the 3D appearance. As w
e track a person, we collect 3D observations over time in a tracklet representat
ion. Given the 3D nature of our observations, we build temporal models for each
one of the previous attributes. We use these models to predict the future state
of the tracklet, including 3D appearance, 3D location, and 3D pose. For a future
 frame, we compute the similarity between the predicted state of a tracklet and
the single frame observations in a probabilistic manner. Association is solved w
ith simple Hungarian matching, and the matches are used to update the respective
 tracklets. We evaluate our approach on various benchmarks and report state-of-t
he-art results. Code and models are available at: https://brjathu.github.io/PHAL
P.
****

Text2Mesh: Text-Driven Neural Stylization for Meshes

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, Rana Hanocka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13492-13502
In this work, we develop intuitive controls for editing the style of 3D objects. Our framework, Text2Mesh, stylizes a 3D mesh by predicting color and local geometric details which conform to a target text prompt. We consider a disentangled representation of a 3D object using a fixed mesh input (content) coupled with a learned neural network, which we term a neural style field network (NSF). In order to modify style, we obtain a similarity score between a text prompt (describing style) and a stylized mesh by harnessing the representational power of CLIP. Text2Mesh requires neither a pre-trained generative model nor a specialized 3D mesh dataset. It can handle low-quality meshes (non-manifold, boundaries, etc.) with arbitrary genus, and does not require UV parameterization. We demonstrate the ability of our technique to synthesize a myriad of styles over a wide variety of 3D meshes. Our code and results are available in our project webpage: https://threedle.github.io/text2mesh/.
****

Learning To Solve Hard Minimal Problems
Petr Hruby, Timothy Duff, Anton Leykin, Tomas Pajdla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5532-5542
We present an approach to solving hard geometric optimization problems in the RANSAC framework. The hard minimal problems arise from relaxing the original geometric optimization problem into a minimal problem with many spurious solutions. Our approach avoids computing large numbers of spurious solutions. We design a learning strategy for selecting a starting problem-solution pair that can be numerically continued to the problem and the solution of interest. We demonstrate our approach by developing a RANSAC solver for the problem of computing the relative pose of three calibrated cameras, via a minimal relaxation using four points in each view. On average, we can solve a single problem in under 70 microseconds. We also benchmark and study our engineering choices on the very familiar problem of computing the relative pose of two calibrated cameras, via the minimal case of five points in two views.
****

H4D: Human 4D Modeling by Learning Neural Compositional Representation
Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19355-19365
Despite the impressive results achieved by deep learning based 3D reconstruction, the techniques of directly learning to model 4D human captures with detailed geometry have been less studied. This work presents a novel framework that can effectively learn a compact and compositional representation for dynamic human by exploiting the human body prior from the widely used SMPL parametric model. Particularly, our representation, named H4D, represents a dynamic 3D human over a temporal span with the SMPL parameters of shape and initial pose, and latent codes encoding motion and auxiliary information. A simple yet effective linear motion model is proposed to provide a rough and regularized motion estimation, followed by per-frame compensation for pose and geometry details with the residual encoded in the auxiliary code. Technically, we introduce novel GRU-based architectures to facilitate learning and improve the representation capability. Extensive experiments demonstrate our method is not only efficacy in recovering dynamic human with accurate motion and detailed geometry, but also amenable to various 4D human related tasks, including motion retargeting, motion completion and future prediction.
****

FWD: Real-Time Novel View Synthesis With Forward Warping and Depth
Ang Cao, Chris Rockwell, Justin Johnson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15713-15724
Novel view synthesis (NVS) is a challenging task requiring systems to generate photorealistic images of scenes from new viewpoints, where both quality and speed

are important for applications. Previous image-based rendering (IBR) methods are fast, but have poor quality when input views are sparse. Recent Neural Radiance Fields (NeRF) and generalizable variants give impressive results but are not real-time. In our paper, we propose a generalizable NVS method with sparse inputs, called \FWDds, which gives high-quality synthesis in real-time. With explicit depth and differentiable rendering, it achieves competitive results to the SOTA methods with 130-1000xspeedup and better perceptual quality. If available, we can seamlessly integrate sensor depth during either training or inference to improve image quality while retaining real-time speed. With the growing prevalence of depths sensors, we hope that methods making use of depth will become increasingly useful.

****

Non-Generative Generalized Zero-Shot Learning via Task-Correlated Disentanglement and Controllable Samples Synthesis

Yaogong Feng, Xiaowen Huang, Pengbo Yang, Jian Yu, Jitao Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9346-9355

Synthesizing pseudo samples is currently the most effective way to solve the Generalized Zero Shot Learning (GZSL) problem. Most models achieve competitive performance but still suffer from two problems: (1) Feature confounding, the overall representations confound task-correlated and task-independent features, and existing models disentangle them in a generative way, but they are unreasonable to synthesize reliable pseudo samples with limited samples; (2) Distribution uncertainty, that massive data is needed when existing models synthesize samples from the uncertain distribution, which causes poor performance in limited samples of seen classes. In this paper, we propose a non-generative model to address these problems correspondingly in two modules: (1) Task-correlated feature disentanglement, to exclude the task-correlated features from task-independent ones by adversarial learning of domain adaption towards reasonable synthesis; (2) Controllable pseudo sample synthesis, to synthesize edge-pseudo and center-pseudo samples with certain characteristics towards more diversity generated and intuitive transfer. In addation, to describe the new scene that is the limit seen class samples in the training process, we further formulate a new ZSL task named the 'Few-shot Seen class and Zero-shot Unseen class learning' (FSZU). Extensive experiments on four benchmarks verify that the proposed method is competitive in the GZSL and the FSZU tasks.

****

C-CAM: Causal CAM for Weakly Supervised Semantic Segmentation on Medical Image

Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, Shaoyi Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11676-11685

Recently, many excellent weakly supervised semantic segmentation (WSSS) works are proposed based on class activation mapping (CAM). However, there are few works that consider the characteristics of medical images. In this paper, we find that there are mainly two challenges of medical images in WSSS: i) the boundary of object foreground and background is not clear; ii) the co-occurrence phenomenon is very severe in training stage. We thus propose a Causal CAM (C-CAM) method to overcome the above challenges. Our method is motivated by two cause-effect chains including category-causality chain and anatomy-causality chain. The category-causality chain represents the image content (cause) affects the category (effect). The anatomy-causality chain represents the anatomical structure (cause) affects the organ segmentation (effect). Extensive experiments were conducted on three public medical image data sets. Our C-CAM generates the best pseudo masks with the DSC of 77.26%, 80.34% and 78.15% on ProMRI, ACDC and CHAOS compared with other CAM-like methods. The pseudo masks of C-CAM are further used to improve the segmentation performance for organ segmentation tasks. Our C-CAM achieves DSC of 83.83% on ProMRI and DSC of 87.54% on ACDC, which outperforms state-of-the-art WSSS methods. Our code is available at https://github.com/Tian-lab/C-CAM.

****

Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, Maja Pantic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14950-14962

One of the most pressing challenges for the detection of face-manipulated videos is generalising to forgery methods not seen during training while remaining effective under common corruptions such as compression. In this paper, we examine whether we can tackle this issue by harnessing videos of real talking faces, which contain rich information on natural facial appearance and behaviour and are readily available in large quantities online. Our method, termed RealForensics, consists of two stages. First, we exploit the natural correspondence between the visual and auditory modalities in real videos to learn, in a self-supervised cross-modal manner, temporally dense video representations that capture factors such as facial movements, expression, and identity. Second, we use these learned representations as targets to be predicted by our forgery detector along with the usual binary forgery classification task; this encourages it to base its real/fake decision on said factors. We show that our method achieves state-of-the-art performance on cross-manipulation generalisation and robustness experiments, and examine the factors that contribute to its performance. Our results suggest that leveraging natural and unlabelled videos is a promising direction for the development of more robust face forgery detectors.

****

Forward Compatible Few-Shot Class-Incremental Learning

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, De-Chuan Zhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9046-9056

Novel classes frequently arise in our dynamically changing world, e.g., new users in the authentication system, and a machine learning model should recognize new classes without forgetting old ones. This scenario becomes more challenging when new class instances are insufficient, which is called few-shot class-incremental learning (FSCIL). Current methods handle incremental learning retrospectively by making the updated model similar to the old one. By contrast, we suggest learning prospectively to prepare for future updates, and propose ForwArd Compatible Training (FACT) for FSCIL. Forward compatibility requires future new classes to be easily incorporated into the current model based on the current stage data, and we seek to realize it by reserving embedding space for future new classes. In detail, we assign virtual prototypes to squeeze the embedding of known classes and reserve for new ones. Besides, we forecast possible new classes and prepare for the updating process. The virtual prototypes allow the model to accept possible updates in the future, which act as proxies scattered among embedding space to build a stronger classifier during inference. FACT efficiently incorporates new classes with forward compatibility and meanwhile resists forgetting of old ones. Extensive experiments validate FACT's state-of-the-art performance. Code is available at: https://github.com/zhoudw-zdw/CVPR22-Fact

****

BaLeNAS: Differentiable Architecture Search via the Bayesian Learning Rule

Miao Zhang, Shirui Pan, Xiaojun Chang, Steven Su, Jilin Hu, Gholamreza (Reza) Haffari, Bin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11871-11880

Differentiable Architecture Search (DARTS) has received massive attention in recent years, mainly because it significantly reduces the computational cost through weight sharing and continuous relaxation. However, more recent works find that existing differentiable NAS techniques struggle to outperform naive baselines, yielding deteriorative architectures as the search proceeds. Rather than directly optimizing the architecture parameters, this paper formulates the neural architecture search as a distribution learning problem through relaxing the architecture weights into Gaussian distributions. By leveraging the natural-gradient variational inference (NGVI), the architecture distribution can be easily optimized based on existing codebases without incurring more memory and computational consumption. We demonstrate how the differentiable NAS benefits from Bayesian principles, enhancing exploration and improving stability. The experimental results on

NAS benchmark datasets confirm the significant improvements the proposed framework can make. In addition, instead of simply applying the argmax on the learned parameters, we further leverage the recently-proposed training-free proxies in NAS to select the optimal architecture from a group architectures drawn from the optimized distribution, where we achieve state-of-the-art results on the NAS-Bench-201 and NAS-Bench-1shot1 benchmarks. Our best architecture in the DARTS search space also obtains competitive test errors with 2.37%, 15.72%, and 24.2% on CIFAR-10, CIFAR-100, and ImageNet, respectively.

****

Cannot See the Forest for the Trees: Aggregating Multiple Viewpoints To Better Classify Objects in Videos

Sukjun Hwang, Miran Heo, Seoung Wug Oh, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17052-17061

Recently, both long-tailed recognition and object tracking have made great advances individually. TAO benchmark presented a mixture of the two, long-tailed object tracking, in order to further reflect the aspect of the real-world. To date, existing solutions have adopted detectors showing robustness in long-tailed distributions, which derive per-frame results. Then, they used tracking algorithms that combine the temporally independent detections to finalize tracklets. However, as the approaches did not take temporal changes in scenes into account, inconsistent classification results in videos led to low overall performance. In this paper, we present a set classifier that improves accuracy of classifying tracklets by aggregating information from multiple viewpoints contained in a tracklet. To cope with sparse annotations in videos, we further propose augmentation of tracklets that can maximize data efficiency. The set classifier is plug-and-playable to existing object trackers, and highly improves the performance of long-tailed object tracking. By simply attaching our method to QDTrack on top of ResNet-101, we achieve the new state-of-the-art, 19.9% and 15.7% TrackAP50 on TAO validation and test sets, respectively.

****

Learning Canonical F-Correlation Projection for Compact Multiview Representation

Yun-Hao Yuan, Jin Li, Yun Li, Jipeng Qiang, Yi Zhu, Xiaobo Shen, Jianping Gou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19260-19269

Canonical correlation analysis (CCA) matters in multiview representation learning. But, CCA and its most variants are essentially based on explicit or implicit covariance matrices. It means that they have no ability to model the nonlinear relationship among features due to intrinsic linearity of covariance. In this paper, we address the preceding problem and propose a novel canonical F-correlation framework by exploring and exploiting the nonlinear relationship between different features. The framework projects each feature rather than observation into a certain new space by an arbitrary nonlinear mapping, thus resulting in more flexibility in real applications. With this framework as a tool, we propose a correlative covariation projection (CCP) method by using an explicit nonlinear mapping. Moreover, we further propose a multiset version of CCP dubbed MCCP for learning compact representation of more than two views. The proposed MCCP is solved by an iterative method, and we prove the convergence of this iteration. A series of experimental results on six benchmark datasets demonstrate the effectiveness of our proposed CCP and MCCP methods.

****

DIFNet: Boosting Visual Information Flow for Image Captioning

Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18020-18029

Current Image captioning (IC) methods predict textual words sequentially based on the input visual information from the visual feature extractor and the partially generated sentence information. However, for most cases, the partially generated sentence may dominate the target word prediction due to the insufficiency of visual information, making the generated descriptions irrelevant to the content

of the given image. In this paper, we propose a Dual Information Flow Network (DIFNet) to address this issue, which takes segmentation feature as another visual information source to enhance the contribution of visual information for prediction. To maximize the use of two information flows, we also propose an effective feature fusion module termed Iterative Independent Layer Normalization (IILN) which can condense the most relevant inputs while retraining modality-specific information in each flow. Experiments show that our method is able to enhance the dependence of prediction on visual information, making word prediction more focused on the visual content, and thus achieve new state-of-the-art performance on the MSCOCO dataset, e.g., 136.2 CIDEr on COCO Karpathy test split.
****

Weakly Supervised Object Localization As Domain Adaption
Lei Zhu, Qi She, Qian Chen, Yunfei You, Boyu Wang, Yanye Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14637-14646
Weakly supervised object localization (WSOL) focuses on localizing objects only with the supervision of image-level classification masks. Most previous WSOL methods follow the classification activation map (CAM) that localizes objects based on the classification structure with the multi-instance learning (MIL) mechanism. However, the MIL mechanism makes CAM only activate discriminative object parts rather than the whole object, weakening its performance for localizing objects. To avoid this problem, this work provides a novel perspective that models WSOL as a domain adaption (DA) task, where the score estimator trained on the source/image domain is tested on the target/pixel domain to locate objects. Under this perspective, a DA-WSOL pipeline is designed to better engage DA approaches into WSOL to enhance localization performance. It utilizes a proposed target sampling strategy to select different types of target samples. Based on these types of target samples, domain adaption localization (DAL) loss is elaborated. It aligns the feature distribution between the two domains by DA and makes the estimator perceive target domain cues by Universum regularization. Experiments show that our pipeline outperforms SOTA methods on multi benchmarks. Code are released at https://github.com/zh460045050/DA-WSOL_CVPR2022.
****

Tencent-MVSE: A Large-Scale Benchmark Dataset for Multi-Modal Video Similarity Evaluation
Zhaoyang Zeng, Yongsheng Luo, Zhenhua Liu, Fengyun Rao, Dian Li, Weidong Guo, Zhen Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3138-3147
Multi-modal video similarity evaluation is important for video recommendation systems such as video de-duplication, relevance matching, ranking, and diversity control. However, there still lacks a benchmark dataset that can support supervised training and accurate evaluation. In this paper, we propose the Tencent-MVSE dataset, which is the first benchmark dataset for the multi-modal video similarity evaluation task. The Tencent-MVSE dataset contains video pairs similarity annotations, and diverse metadata including Chinese title, automatic speech recognition (ASR) text, as well as human-annotated categories/tags. We provide a simple baseline with a multi-modal Transformer architecture to perform supervised multi-modal video similarity evaluation. We also explore pre-training strategies to make use of the unpaired data. The whole dataset as well as our baseline will be released to promote the development of the multi-modal video similarity evaluation. The dataset has been released in https://tencent-mvse.github.io/.
****

Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation
Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, Efstratios Gavves; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11553-11562
The key challenge for few-shot semantic segmentation (FSS) is how to tailor a desirable interaction among support and query features and/or their prototypes, under the episodic training scenario. Most existing FSS methods implement such support/query interactions by solely leveraging \it plain operations -- e.g., cosi

ne similarity and feature concatenation -- for segmenting the query objects. How ever, these interaction approaches usually cannot well capture the intrinsic obj ect details in the query images that are widely encountered in FSS, e.g., if the query object to be segmented has holes and slots, inaccurate segmentation almos t always happens. To this end, we propose a dynamic prototype convolution networ k (DPCN) to fully capture the aforementioned intrinsic details for accurate FSS. Specifically, in DPCN, a dynamic convolution module (DCM) is firstly proposed t o generate dynamic kernels from support foreground, then information interaction is achieved by convolution operations over query features using these kernels. Moreover, we equip DPCN with a support activation module (SAM) and a feature fil tering module (FFM) to generate pseudo mask and filter out background informatio n for the query images, respectively. SAM and FFM together can mine enriched con text information from the query features. Our DPCN is also flexible and efficien t under the k-shot FSS setting. Extensive experiments on PASCAL-5^i and COCO-20^ i show that DPCN yields superior performances under both 1-shot and 5-shot setti ngs.

****

Deep Orientation-Aware Functional Maps: Tackling Symmetry Issues in Shape Matchi ng

Nicolas Donati, Etienne Corman, Maks Ovsjanikov; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 742-751

State-of-the-art fully intrinsic network for non-rigid shape matching are unable to disambiguate between shape inner symmetries. Meanwhile, recent advances in t he functional map framework allow to enforce orientation preservation using a fu nctional representation for tangent vector field transfer, through so-called com plex functional maps. Using this representation, we propose a new deep learning approach to learn orientation-aware features in a fully unsupervised setting. Ou r architecture is built on DiffusionNet, which makes our method robust to discre tization changes, while adding a vector-field-based loss, which promotes orienta tion preservation without using (often unstable) extrinsic descriptors. Our sour ce code is available at: https://github.com/nicolasdonati/DUO-FM

****

Tree Energy Loss: Towards Sparsely Annotated Semantic Segmentation

Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 022, pp. 16907-16916

Sparsely annotated semantic segmentation (SASS) aims to train a segmentation net work with coarse-grained (i.e.,point-, scribble-, and block-wise) supervisions, where only a small proportion of pixels are labeled in each image. In this paper , we propose a novel tree energy loss for SASS by providing semantic guidance fo r unlabeled pixels. The tree energy loss represents images as minimum spanning t rees to model both low-level and high-level pair-wise affinities. By sequentiall y applying these affinities to the network prediction, soft pseudo labels for un labeled pixels are generated in a coarse-to-fine manner, resulting in dynamic on line self-training. The tree energy loss is effective and easy to be incorporate d into existing frameworks by combining it with a traditional segmentation loss. Compared with previous SASS methods, our method requires no multi-stage trainin g strategies, alternating optimization procedures, additional supervised data, o r time-consuming post-processing while outperforming them in all types of superv ised settings. Code is available at https://github.com/megvii-research/TreeEnerg yLoss.

****

Mr.BiQ: Post-Training Non-Uniform Quantization Based on Minimizing the Reconstru ction Error

Yongkweon Jeon, Chungman Lee, Eulrang Cho, Yeonju Ro; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12329- 12338

Post-training quantization compresses a neural network within few hours with onl y a small unlabeled calibration set. However, so far it has been only discussed and empirically demonstrated in the context of uniform quantization on convoluti

onal neural networks. We thus propose a new post-training non-uniform quantizati on method, called Mr.BiQ, allowing low bit-width quantization even on Transforme r models. In particular, we leverage multi-level binarization for weights while allowing activations to be represented as various data formats (e.g., INT8, bflo at16, binary-coding, and FP32). Unlike conventional methods which optimize full-precision weights first, then decompose the weights into quantization parameters , Mr.BiQ recognizes the quantization parameters (i.e., scaling factors and bit-c ode) as directly and jointly learnable parameters during the optimization. To ve rify the superiority of the proposed quantization scheme, we test Mr.BiQ on vari ous models including convolutional neural networks and Transformer models. Accor ding to experimental results, Mr.BiQ shows significant improvement in terms of a ccuracy when the bit-width of weights is equal to 2: up to 5.35 p.p. improvement in CNNs, up to 4.23 p.p. improvement in Vision Transformers, and up to 3.37 poi nt improvement in Transformers for NLP.

****

MatteFormer: Transformer-Based Image Matting via Prior-Tokens

GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, Nojun Kwak; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 11696-11706

In this paper, we propose a transformer-based image matting model called MatteFo rmer, which takes full advantage of trimap information in the transformer block. Our method first introduces a prior-token which is a global representation of e ach trimap region (e.g. foreground, background and unknown). These prior-tokens are used as global priors and participate in the self-attention mechanism of eac h block. Each stage of the encoder is composed of PAST (Prior-Attentive Swin Tra nsformer) block, which is based on the Swin Transformer block, but differs in a couple of aspects: 1) It has PA-WSA (Prior-Attentive Window Self-Attention) laye r, performing self-attention not only with spatial-tokens but also with prior-to kens. 2) It has prior-memory which saves prior-tokens accumulatively from the pr evious blocks and transfers them to the next block. We evaluate our MatteFormer on the commonly used image matting datasets: Composition-1k and Distinctions-646 . Experiment results show that our proposed method achieves state-of-the-art per formance with a large margin. Our codes are available at https://github.com/webt oon/matteformer.

****

Video Shadow Detection via Spatio-Temporal Interpolation Consistency Training

Xiao Lu, Yihong Cao, Sheng Liu, Chengjiang Long, Zipei Chen, Xuanyu Zhou, Yimin Yang, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 3116-3125

It is challenging to annotate large-scale datasets for supervised video shadow d etection methods. Using a model trained on labeled images to the video frames di rectly may lead to high generalization error and temporal inconsistent results. In this paper, we address these challenges by proposing a Spatio-Temporal Interp olation Consistency Training (STICT) framework to rationally feed the unlabeled video frames together with the labeled images into an image shadow detection net work training. Specifically, we propose the Spatial and Temporal ICT, in which w e define two new interpolation schemes, i.e., the spatial interpolation and the temporal interpolation. We then derive the spatial and temporal interpolation co nsistency constraints accordingly for enhancing generalization in the pixel-wise classification task and for encouraging temporal consistent predictions, respec tively. In addition, we design a Scale-Aware Network for multi-scale shadow know ledge learning in images, and propose a scale-consistency constraint to minimize the discrepancy among the predictions at different scales. Our proposed approac h is extensively validated on the ViSha dataset and a self-annotated dataset. Ex perimental results show that, even without video labels, our approach is better than most state of the art supervised, semi-supervised or unsupervised image/vid eo shadow detection methods and other methods in related tasks. Code and dataset are available at https://github.com/yihong-97/STICT.

****

Ranking Distance Calibration for Cross-Domain Few-Shot Learning

Pan Li, Shaogang Gong, Chengjie Wang, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9099-9108
Recent progress in few-shot learning promotes a more realistic cross-domain setting, where the source and target datasets are in different domains. Due to the domain gap and disjoint label spaces between source and target datasets, their shared knowledge is extremely limited. This encourages us to explore more information in the target domain rather than to overly elaborate training strategies on the source domain as in many existing methods. Hence, we start from a generic representation pre-trained by a cross-entropy loss and a conventional distance-based classifier, along with an image retrieval view, to employ a re-ranking process to calibrate a target distance matrix by discovering the k-reciprocal neighbours within the task. Assuming the pre-trained representation is biased towards the source, we construct a non-linear subspace to minimise task-irrelevant features therewithin while keep more transferrable discriminative information by a hyperbolic tangent transformation. The calibrated distance in this target-aware non-linear sub-space is complementary to that in the pre-trained representation. To impose such distance calibration information onto the pre-trained representation, a Kullback-Leibler divergence loss is employed to gradually guide the model towards the calibrated distance-based distribution. Extensive evaluations on eight target domains show that this target ranking calibration process can improve conventional distance-based classifiers in few-shot learning.
****

Robust and Accurate Superquadric Recovery: A Probabilistic Approach
Weixiao Liu, Yuwei Wu, Sipu Ruan, Gregory S. Chirikjian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2676-2685
Interpreting objects with basic geometric primitives has long been studied in computer vision. Among geometric primitives, superquadrics are well known for their ability to represent a wide range of shapes with few parameters. However, as the first and foremost step, recovering superquadrics accurately and robustly from 3D data still remains challenging. The existing methods are subject to local optima and sensitive to noise and outliers in real-world scenarios, resulting in frequent failure in capturing geometric shapes. In this paper, we propose the first probabilistic method to recover superquadrics from point clouds. Our method builds a Gaussian-uniform mixture model (GUM) on the parametric surface of a superquadric, which explicitly models the generation of outliers and noise. The superquadric recovery is formulated as a Maximum Likelihood Estimation (MLE) problem. We propose an algorithm, Expectation, Maximization, and Switching (EMS), to solve this problem, where: (1) outliers are predicted from the posterior perspective; (2) the superquadric parameter is optimized by the trust-region reflective algorithm; and (3) local optima are avoided by globally searching and switching among parameters encoding similar superquadrics. We show that our method can be extended to the multi-superquadrics recovery for complex objects. The proposed method outperforms the state-of-the-art in terms of accuracy, efficiency, and robustness on both synthetic and real-world datasets. The code is at http://github.com/bmlklwx/EMS-superquadric_fitting.git.
****

Zero-Shot Text-Guided Object Generation With Dream Fields
Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, Ben Poole; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 867-876
We combine neural rendering with multi-modal image and text representations to synthesize diverse 3D objects solely from natural language descriptions. Our method, Dream Fields, can generate the geometry and color of a wide range of objects without 3D supervision. Due to the scarcity of diverse, captioned 3D data, prior methods only generate objects from a handful of categories, such as ShapeNet. Instead, we guide generation with image-text models pre-trained on large datasets of captioned images from the web. Our method optimizes a Neural Radiance Field from many camera views so that rendered images score highly with a target caption according to a pre-trained CLIP model. To improve fidelity and visual quality

, we introduce simple geometric priors, including sparsityinducing transmittance regularization, scene bounds, and new MLP architectures. In experiments, Dream Fields produce realistic, multi-view consistent object geometry and color from a variety of natural language captions.
****

Learning Pixel Trajectories With Multiscale Contrastive Random Walks
Zhangxing Bian, Allan Jabri, Alexei A. Efros, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6508-6519
A range of video modeling tasks, from optical flow to multiple object tracking, share the same fundamental challenge: establishing space-time correspondence. Yet, approaches that dominate each space differ. We take a step towards bridging this gap by extending the recent contrastive random walk formulation to much more dense, pixel-level space-time graphs. The main contribution is introducing hierarchy into the search problem by computing the transition matrix in a coarse-to-fine manner, forming a multiscale contrastive random walk. This establishes a unified technique for self-supervised learning of optical flow, keypoint tracking, and video object segmentation. Experiments demonstrate that, for each of these tasks, our unified model achieves performance competitive with strong self-supervised approaches specific to that task.
****

Self-Supervised Correlation Mining Network for Person Image Generation
Zijian Wang, Xingqun Qi, Kun Yuan, Muyi Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7703-7712
Person image generation aims to perform non-rigid deformation on source images, which generally requires unaligned data pairs for training. Recently, self-supervised methods express great prospects in this task by merging the disentangled representations for self-reconstruction. However, such methods fail to exploit the spatial correlation between the disentangled features. In this paper, we propose a Self-supervised Correlation Mining Network (SCM-Net) to rearrange the source images in the feature space, in which two collaborative modules are integrated, Decomposed Style Encoder (DSE) and Correlation Mining Module (CMM). Specifically, the DSE first creates unaligned pairs at the feature level. Then, the CMM establishes the spatial correlation field for feature rearrangement. Eventually, a translation module transforms the rearranged features to realistic results. Meanwhile, for improving the fidelity of cross-scale pose transformation, we propose a graph based Body Structure Retaining Loss (BSR Loss) to preserve reasonable body structures on half body to full body generation. Extensive experiments conducted on DeepFashion dataset demonstrate the superiority of our method compared with other supervised and unsupervised approaches. Furthermore, satisfactory results on face generation show the versatility of our method in other deformation tasks.
****

Grounding Answers for Visual Questions Asked by Visually Impaired People
Chongyan Chen, Samreen Anjum, Danna Gurari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19098-19107
Visual question answering is the task of answering questions about images. We introduce the VizWiz-VQA-Grounding dataset, the first dataset that visually grounds answers to visual questions asked by people with visual impairments. We analyze our dataset and compare it with five VQA-Grounding datasets to demonstrate what makes it similar and different. We then evaluate the SOTA VQA and VQA-Grounding models and demonstrate that current SOTA algorithms often fail to identify the correct visual evidence where the answer is located. These models regularly struggle when the visual evidence occupies a small fraction of the image, for images that are higher quality, as well as for visual questions that require skills in text recognition. The dataset, evaluation server, and leaderboard all can be found at the following link: https://vizwiz.org/tasks-and-datasets/answer-grounding-for-vqa/.
****

Task Adaptive Parameter Sharing for Multi-Task Learning

Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7561-7570
Adapting pre-trained models with broad capabilities has become standard practice for learning a wide range of downstream tasks. The typical approach of fine-tuning different models for each task is performant, but incurs a substantial memory cost. To efficiently learn multiple downstream tasks we introduce Task Adaptive Parameter Sharing (TAPS), a simple method for tuning a base model to a new task by adaptively modifying a small, task-specific subset of layers. This enables multi-task learning while minimizing the resources used and avoids catastrophic forgetting and competition between tasks. TAPS solves a joint optimization problem which determines both the layers that are shared with the base model and the value of the task-specific weights. Further, a sparsity penalty on the number of active layers promotes weight sharing with the base model. Compared to other methods, TAPS retains a high accuracy on the target tasks while still introducing only a small number of task-specific parameters. Moreover, TAPS is agnostic to the particular architecture used and requires only minor changes to the training scheme. We evaluate our method on a suite of fine-tuning tasks and architectures (ResNet, DenseNet, ViT) and show that it achieves state-of-the-art performance while being simple to implement.
****
Sparse Instance Activation for Real-Time Instance Segmentation
Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4433-4442
In this paper, we propose a conceptually novel, efficient, and fully convolutional framework for real-time instance segmentation. Previously, most instance segmentation methods heavily rely on object detection and perform mask prediction based on bounding boxes or dense centers. In contrast, we propose a sparse set of instance activation maps, as a new object representation, to highlight informative regions for each foreground object. Then instance-level features are obtained by aggregating features according to the highlighted regions for recognition and segmentation. Moreover, based on bipartite matching, the instance activation maps can predict objects in a one-to-one style, thus avoiding non-maximum suppression (NMS) in post-processing. Owing to the simple yet effective designs with instance activation maps, SparseInst has extremely fast inference speed and achieves 40 FPS and 37.9 AP on the COCO benchmark, which significantly outperforms the counterparts in terms of speed and accuracy. Code and models are available at https://github.com/hustvl/SparseInst.
****
Automatic Color Image Stitching Using Quaternion Rank-1 Alignment
Jiaxue Li, Yicong Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19720-19729
Color image stitching is a challenging task in real-world applications. This paper first proposes a quaternion rank-1 alignment (QR1A) model for high-precision color image alignment. To solve the optimization problem of QR1A, we develop a nested iterative algorithm under the framework of complex-valued alternating direction method of multipliers. To quantitatively evaluate image stitching performance, we propose a perceptual seam quality (PSQ) measure to calculate misalignments of local regions along the seamline. Using QR1A and PSQ, we further propose an automatic color image stitching (ACIS-QR1A) framework. In this framework, the automatic strategy and iterative learning strategy are developed to simultaneously learn the optimal seamline and local alignment. Extensive experiments on challenging datasets demonstrate that the proposed ACIS-QR1A is able to obtain high-quality stitched images under several difficult scenarios including large parallax, low textures, moving objects, large occlusions or/and their combinations.
****
VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning
Jun Chen, Han Guo, Kai Yi, Boyang Li, Mohamed Elhoseiny; Proceedings of the IEEE

/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18030-18040
The limited availability of annotated data often hinders real-world applications of machine learning. To efficiently learn from small quantities of multimodal data, we leverage the linguistic knowledge from a large pre-trained language model (PLM) and quickly adapt it to new domains of image captioning. To effectively utilize a pretrained model, it is critical to balance the visual input and prior linguistic knowledge from pretraining. We propose VisualGPT, which employs a novel self-resurrecting encoder-decoder attention mechanism to quickly adapt the PLM with a small amount of in-domain image-text data. The proposed self-resurrecting activation unit produces sparse activations that prevent accidental overwriting of linguistic knowledge. When trained on 0.1%, 0.5% and 1% of the respective training sets, VisualGPT surpasses the best baseline by up to 10.0% CIDEr on MS COCO and 17.9% CIDEr on Conceptual Captions. Furthermore, VisualGPT achieves the state-of-the-art result on IU X-ray, a medical report generation dataset. Our code is available at https://github.com/Vision-CAIR/VisualGPT.
****

ESCNet: Gaze Target Detection With the Understanding of 3D Scenes
Jun Bao, Buyu Liu, Jun Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14126-14135
This paper aims to address the single image gaze target detection problem. Conventional methods either focus on 2D visual cues or exploit additional depth information in a very coarse manner. In this work, we propose to explicitly and effectively model 3D geometry under challenging scenario where only 2D annotations are available. We first obtain 3D point clouds of given scene with estimated depth and reference objects. Then we figure out the front-most points in all possible 3D directions of given person. These points are later leveraged in our ESCNet model. Specifically, ESCNet consists of geometry and scene parsing modules. The former produces an initial heatmap inferring the probability that each front-most point has been looking at according to estimated 3D gaze direction. And the latter further explores scene contextual cues to regulate detection results. We validate our idea on two publicly available dataset, GazeFollow and VideoAttentionTarget, and demonstrate the state-of-the-art performance. Our method also beats the human in terms of AUC on GazeFollow.
****

Can You Spot the Chameleon? Adversarially Camouflaging Images From Co-Salient Object Detection
Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Huazhu Fu, Wei Feng, Yang Liu, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2150-2159
Co-salient object detection (CoSOD) has recently achieved significant progress and played a key role in retrieval-related tasks. However, it inevitably poses an entirely new safety and security issue, i.e., highly personal and sensitive content can potentially be extracting by powerful CoSOD methods. In this paper, we address this problem from the perspective of adversarial attacks and identify a novel task: adversarial co-saliency attack. Specially, given an image selected from a group of images containing some common and salient objects, we aim to generate an adversarial version that can mislead CoSOD methods to predict incorrect co-salient regions. Note that, compared with general white-box adversarial attacks for classification, this new task faces two additional challenges: (1) low success rate due to the diverse appearance of images in the group; (2) low transferability across CoSOD methods due to the considerable difference between CoSOD pipelines. To address these challenges, we propose the very first black-box joint adversarial exposure and noise attack (Jadena), where we jointly and locally tune the exposure and additive perturbations of the image according to a newly designed high-feature-level contrast-sensitive loss function. Our method, without any information on the state-of-the-art CoSOD methods, leads to significant performance degradation on various co-saliency detection datasets and makes the co-salient objects undetectable. This can have strong practical benefits in properly securing the large number of personal photos currently shared on the Internet. M

oreover, our method is potential to be utilized as a metric for evaluating the r
obustness of CoSOD methods.
****

Finding Badly Drawn Bunnies
Lan Yang, Kaiyue Pang, Honggang Zhang, Yi-Zhe Song; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7482-749
1
As lovely as bunnies are, your sketched version would probably not do it justice
 (Fig. 1). This paper recognises this very problem and studies sketch quality me
asurement for the first time -- letting you find these badly drawn ones. Our key
 discovery lies in exploiting the magnitude (L2 norm) of a sketch feature as a q
uantitative quality metric. We propose Geometry-Aware Classification Layer (GACL
), a generic method that makes feature-magnitude-as-quality-metric possible and
importantly does it without the need for specific quality annotations from human
s. GACL sees feature magnitude and recognisability learning as a dual task, whic
h can be simultaneously optimised under a neat cross-entropy classification loss
. GACL is lightweight with theoretic guarantees and enjoys a nice geometric inte
rpretation to reason its success. We confirm consistent quality agreements betwe
en our GACL-induced metric and human perception through a carefully designed hum
an study. Last but not least, we demonstrate three practical sketch applications
 enabled for the first time using our quantitative quality metric. Code will be
made publicly available.
****

Point2Cyl: Reverse Engineering 3D Objects From Point Clouds to Extrusion Cylinde
rs
Mikaela Angelina Uy, Yen-Yu Chang, Minhyuk Sung, Purvi Goel, Joseph G. Lambourne
, Tolga Birdal, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 11850-11860
We propose Point2Cyl, a supervised network transforming a raw 3D point cloud to
a set of extrusion cylinders. Reverse engineering from a raw geometry to a CAD m
odel is an essential task to enable manipulation of the 3D data in shape editing
 software and thus expand their usages in many downstream applications. Particul
arly, the form of CAD models having a sequence of extrusion cylinders --- a 2D s
ketch plus an extrusion axis and range --- and their boolean combinations is not
 only widely used in the CAD community/software but also has great expressivity
of shapes, compared to having limited types of primitives (e.g., planes, spheres
, and cylinders). In this work, we introduce a neural network that solves the ex
trusion cylinder decomposition problem in a geometry-grounded way by first learn
ing underlying geometric proxies. Precisely, our approach first predicts per-poi
nt segmentation, base/barrel labels and normals, then estimates for the underlyi
ng extrusion parameters in differentiable and closed-form formulations. Our expe
riments show that our approach demonstrates the best performance on two recent C
AD datasets, Fusion Gallery and DeepCAD, and we further showcase our approach on
 reverse engineering and editing.
****

All-Photon Polarimetric Time-of-Flight Imaging
Seung-Hwan Baek, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2022, pp. 17876-17885
Time-of-flight (ToF) sensors provide an image modality fueling applications acro
ss domains, including lidar in autonomous driving, robotics, and augmented reali
ty. Conventional ToF imaging methods estimate the depth of a scene point by send
ing pulses of light into a scene and measuring the time of flight of the first a
rriving photons that are returned from the scene, the ones directly reflected fr
om a scene surface without any temporal delay. As such, all photons following th
is first response are typically considered as unwanted noise, including multi-bo
unce and sub-surface scattering of real-world materials. While multi-bounce scen
e interreflections have been extensively in recent work on non-line-of-sight ima
ging, we investigate temporally resolved sub-surface scattering in this work. We
 depart from the principle of first arrival and instead propose an all-photon To
F imaging method relying on polarization changes that analyzes both first- and l

ate-arriving photons for shape and material scene understanding. To this end, we propose a novel capture method, reflectance model, and a reconstruction algorithm that exploits the polarization state of light changes after reflection in addition to ToF information. The proposed temporal-polarimetric imaging method allows for accurate geometric and material information of the scene by utilizing all photons captured by the system, decoded by polarization cues, outperforming all tested existing methods in simulation and experimentally.
****

MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation
Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13147-13156
Estimating 3D human poses from monocular videos is a challenging task due to depth ambiguity and self-occlusion. Most existing works attempt to solve both issues by exploiting spatial and temporal relationships. However, those works ignore the fact that it is an inverse problem where multiple feasible solutions (i.e., hypotheses) exist. To relieve this limitation, we propose a Multi-Hypothesis Transformer (MHFormer) that learns spatio-temporal representations of multiple plausible pose hypotheses. In order to effectively model multi-hypothesis dependencies and build strong relationships across hypothesis features, the task is decomposed into three stages: (i) Generate multiple initial hypothesis representations; (ii) Model self-hypothesis communication, merge multiple hypotheses into a single converged representation and then partition it into several diverged hypotheses; (iii) Learn cross-hypothesis communication and aggregate the multi-hypothesis features to synthesize the final 3D pose. Through the above processes, the final representation is enhanced and the synthesized pose is much more accurate. Extensive experiments show that MHFormer achieves state-of-the-art results on two challenging datasets: Human3.6M and MPI-INF-3DHP. Without bells and whistles, its performance surpasses the previous best result by a large margin of 3% on Human3.6M. Code and models are available at https://github.com/Vegetebird/MHFormer.
****

Surface-Aligned Neural Radiance Fields for Controllable 3D Human Synthesis
Tianhan Xu, Yasuhiro Fujita, Eiichi Matsumoto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15883-15892
We propose a new method for reconstructing controllable implicit 3D human models from sparse multi-view RGB videos. Our method defines the neural scene representation on the mesh surface points and signed distances from the surface of a human body mesh. We identify an indistinguishability issue that arises when a point in 3D space is mapped to its nearest surface point on a mesh for learning surface-aligned neural scene representation. To address this issue, we propose projecting a point onto a mesh surface using a barycentric interpolation with modified vertex normals. Experiments with the ZJU-MoCap and Human3.6M datasets show that our approach achieves a higher quality in a novel-view and novel-pose synthesis than existing methods. We also demonstrate that our method easily supports the control of body shape and clothes. Project page: https://pfnet-research.github.io/surface-aligned-nerf/.
****

Learning From Temporal Gradient for Semi-Supervised Action Recognition
Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, Yingwei Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3252-3262
Semi-supervised video action recognition tends to enable deep neural networks to achieve remarkable performance even with very limited labeled data. However, existing methods are mainly transferred from current image-based methods (e.g., FixMatch). Without specifically utilizing the temporal dynamics and inherent multimodal attributes, their results could be suboptimal. To better leverage the encoded temporal information in videos, we introduce temporal gradient as an additional modality for more attentive feature extraction in this paper. To be specific, our method explicitly distills the fine-grained motion representations from temporal gradient (TG) and imposes consistency across different modalities (i.e.,

RGB and TG). The performance of semi-supervised action recognition is significantly improved without additional computation or parameters during inference. Our method achieves the state-of-the-art performance on three video action recognition benchmarks (i.e., Kinetics-400, UCF-101, and HMDB-51) under several typical semi-supervised settings (i.e., different ratios of labeled data).
****

Towards Implicit Text-Guided 3D Shape Generation
Zhengzhe Liu, Yi Wang, Xiaojuan Qi, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17896-17906
In this work, we explore the challenging task of generating 3D shapes from text. Beyond the existing works, we propose a new approach for text-guided 3D shape generation, capable of producing high-fidelity shapes with colors that match the given text description. This work has several technical contributions. First, we decouple the shape and color predictions for learning features in both texts and shapes, and propose the word-level spatial transformer to correlate word features from text with spatial features from shape. Also, we design a cyclic loss to encourage consistency between text and shape, and introduce the shape IMLE to diversify the generated shapes. Further, we extend the framework to enable text-guided shape manipulation. Extensive experiments on the largest existing text-shape benchmark manifest the superiority of this work. The code and the models are available at https://github.com/liuzhengzhe/Towards-Implicit Text-Guided-Shape-Generation.
****

Audio-Driven Neural Gesture Reenactment With Video Motion Graphs
Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, Evangelos Kalogerakis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3418-3428
Human speech is often accompanied by body gestures including arm and hand gestures. We present a method that reenacts a high-quality video with gestures matching a target speech audio. The key idea of our method is to split and re-assemble clips from a reference video through a novel video motion graph encoding valid transitions between clips. To seamlessly connect different clips in the reenactment, we propose a pose-aware video blending network which synthesizes video frames around the stitched frames between two clips. Moreover, we developed an audio-based gesture searching algorithm to find the optimal order of the reenacted frames. Our system generates reenactments that are consistent with both the audio rhythms and the speech content. We evaluate our synthesized video quality quantitatively, qualitatively, and with user studies, demonstrating that our method produces videos of much higher quality and consistency with the target audio compared to previous work and baselines.
****

SoftCollage: A Differentiable Probabilistic Tree Generator for Image Collage
Jiahao Yu, Li Chen, Mingrui Zhang, Mading Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3729-3738
Image collage task aims to create an informative and visual-aesthetic visual summarization for an image collection. While several recent works exploit tree-based algorithm to preserve image content better, all of them resort to hand-crafted adjustment rules to optimize the collage tree structure, leading to the failure of fully exploring the structure space of collage tree. Our key idea is to soften the discrete tree structure space into a continuous probability space. We propose SoftCollage, a novel method that employs a neural-based differentiable probabilistic tree generator to produce the probability distribution of correlation-preserving collage tree conditioned on deep image feature, aspect ratio and canvas size. The differentiable characteristic allows us to formulate the tree-based collage generation as a differentiable process and directly exploit gradient to optimize the collage layout in the level of probability space in an end-to-end manner. To facilitate image collage research, we propose AIC, a large-scale public-available annotated dataset for image collage evaluation. Extensive experiments on the introduced dataset demonstrate the superior performance of the proposed method. Data and codes are available at https://github.com/ChineseYjh/SoftColl

age.
****
Transforming Model Prediction for Tracking
Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8731-8740
Optimization based tracking methods have been widely successful by integrating a target model prediction module, providing effective global reasoning by minimizing an objective function. While this inductive bias integrates valuable domain knowledge, it limits the expressivity of the tracking network. In this work, we therefore propose a tracker architecture employing a Transformer-based model prediction module. Transformers capture global relations with little inductive bias, allowing it to learn the prediction of more powerful target models. We further extend the model predictor to estimate a second set of weights that are applied for accurate bounding box regression. The resulting tracker relies on training and on test frame information in order to predict all weights transductively. We train the proposed tracker end-to-end and validate its performance by conducting comprehensive experiments on multiple tracking datasets. Our tracker sets a new state of the art on three benchmarks, achieving an AUC of 68.5% on the challenging LaSOT dataset.
****
A Unified Framework for Implicit Sinkhorn Differentiation
Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Florian Bernard, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 509-518
The Sinkhorn operator has recently experienced a surge of popularity in computer vision and related fields. One major reason is its ease of integration into deep learning frameworks. To allow for an efficient training of respective neural networks, we propose an algorithm that obtains analytical gradients of a Sinkhorn layer via implicit differentiation. In comparison to prior work, our framework is based on the most general formulation of the Sinkhorn operator. It allows for any type of loss function, while both the target capacities and cost matrices are differentiated jointly. We further construct error bounds of the resulting algorithm for approximate inputs. Finally, we demonstrate that for a number of applications, simply replacing automatic differentiation with our algorithm directly improves the stability and accuracy of the obtained gradients. Moreover, we show that it is computationally more efficient, particularly when resources like GPU memory are scarce.
****
DGECN: A Depth-Guided Edge Convolutional Network for End-to-End 6D Pose Estimation
Tuo Cao, Fei Luo, Yanping Fu, Wenxiao Zhang, Shengjie Zheng, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3783-3792
Monocular 6D pose estimation is a fundamental task in computer vision. Existing works often adopt a twostage pipeline by establishing correspondences and utilizing a RANSAC algorithm to calculate 6 degrees-of-freedom (6DoF) pose. Recent works try to integrate differentiable RANSAC algorithms to achieve an end-to-end 6D pose estimation. However, most of them hardly consider the geometric features in 3D space, and ignore the topology cues when performing differentiable RANSAC algorithms. To this end, we proposed a Depth-Guided Edge Convolutional Network (DGECN) for 6D pose estimation task. We have made efforts from the following three aspects: 1) We take advantages of estimated depth information to guide both the correspondences-extraction process and the cascaded differentiable RANSAC algorithm with geometric information. 2)We leverage the uncertainty of the estimated depth map to improve accuracy and robustness of the output 6D pose. 3) We propose a differentiable Perspective-n-Point(PnP) algorithm via edge convolution to explore the topology relations between 2D-3D correspondences. Experiments demonstrate that our proposed network outperforms current works on both effectiveness and efficiency.

****
Unsupervised Vision-Language Parsing: Seamlessly Bridging Visual Scene Graphs Wi
th Language Structures via Dependency Relationships
Chao Lou, Wenjuan Han, Yuhuan Lin, Zilong Zheng; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15607-15616
Understanding realistic visual scene images together with language descriptions
 is a fundamental task towards generic visual understanding. Previous works have
 shown compelling comprehensive results by building hierarchical structures for v
isual scenes (e.g., scene graphs) and natural languages (e.g., dependency trees)
, individually. However, how to construct a joint vision-language (VL) structure
 has barely been investigated. More challenging but worthwhile, we introduce a n
ew task that targets on inducing such a joint VL structure in an unsupervised ma
nner. Our goal is to bridge the visual scene graphs and linguistic dependency tr
ees seamlessly. Due to the lack of VL structural data, we start by building a ne
w dataset VLParse. Rather than using labor-intensive labeling from scratch, we p
ropose an automatic alignment procedure to produce coarse structures followed by
 human refinement to produce high-quality ones. Moreover, we benchmark our datas
et by proposing a contrastive learning (CL)-based framework VLGAE, short for Vis
ion-Language Graph Autoencoder. Our model obtains superior performance on two de
rived tasks, i.e., language grammar induction and VL phrase grounding. Ablations
 show the effectiveness of both visual cues and dependency relationships on fine
-grained VL structure construction.
****
Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling
Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, Ehsan Elhamifar; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 7020-7031
Open-vocabulary instance segmentation aims at segmenting novel classes without m
ask annotations. It is an important step toward reducing laborious human supervi
sion. Most existing works first pretrain a model on captioned images covering ma
ny novel classes and then finetune it on limited base classes with mask annotati
ons. However, the high-level textual information learned from caption pretrainin
g alone cannot effectively encode the details required for pixel-wise segmentati
on. To address this, we propose a cross-modal pseudo-labeling framework, which g
enerates training pseudo masks by aligning word semantics in captions with visua
l features of object masks in images. Thus, our framework is capable of labeling
 novel classes in captions via their word semantics to self-train a student mode
l. To account for noises in pseudo masks, we design a robust student model that
selectively distills mask knowledge by estimating the mask noise levels, hence m
itigating the adverse impact of noisy pseudo masks. By extensive experiments, we
 show the effectiveness of our framework, where we significantly improve mAP sco
re by 4.5% on MS-COCO and 5.1% on the large-scale Open Images & Conceptual Capti
ons datasets compared to the state-of-the-art.
****
Locality-Aware Inter- and Intra-Video Reconstruction for Self-Supervised Corresp
ondence Learning
Liulei Li, Tianfei Zhou, Wenguan Wang, Lu Yang, Jianwu Li, Yi Yang; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 8719-8730
Our target is to learn visual correspondence from unlabeled videos. We develop L
IIR, a locality-aware inter-and intra-video reconstruction framework that fills
in three missing pieces, i.e., instance discrimination, location awareness, and
spatial compactness, of self-supervised correspondence learning puzzle. First, i
nstead of most existing efforts focusing on intra-video self-supervision only, w
e exploit cross video affinities as extra negative samples within a unified, int
er-and intra-video reconstruction scheme. This enables instance discriminative r
epresentation learning by contrasting desired intra-video pixel association agai
nst negative inter-video correspondence. Second, we merge position information i
nto correspondence matching, and design a position shifting strategy to remove t
he side-effect of position encoding during inter-video affinity computation, mak

ing our LIIR location-sensitive. Third, to make full use of the spatial continuity nature of video data, we impose a compactness-based constraint on correspondence matching, yielding more sparse and reliable solutions. The learned representation surpasses self-supervised state-of-the-arts on label propagation tasks including objects, semantic parts, and keypoints.

****

A Versatile Multi-View Framework for LiDAR-Based 3D Object Detection With Guidance From Panoptic Segmentation

Hamidreza Fazlali, Yixuan Xu, Yuan Ren, Bingbing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17192-17201

3D object detection using LiDAR data is an indispensable component for autonomous driving systems. Yet, only a few LiDAR-based 3D object detection methods leverage segmentation information to further guide the detection process. In this paper, we propose a novel multi-task framework that jointly performs 3D object detection and panoptic segmentation. In our method, the 3D object detection backbone, which is in Bird's-Eye-View (BEV) plane, is augmented by the injection of Range-View (RV) feature maps from the 3D panoptic segmentation backbone. This enables the detection backbone to leverage multi-view information to address the shortcomings of each projection view. Furthermore, foreground semantic information is incorporated to ease the detection task by highlighting the locations of each object class in the feature maps. Finally, a new center density heatmap generated based on the instance-level information further guides the detection backbone by suggesting possible box center locations for objects in the BEV plane. Our method works with any BEV-based 3D object detection method, and as shown by extensive experiments on the nuScenes dataset, it provides significant performance gains. Notably, the proposed method based on a single-stage CenterPoint 3D object detection network achieved state-of-the-art performance on nuScenes 3D Detection Benchmark with 67.3 NDS.

****

Query and Attention Augmentation for Knowledge-Based Explainable Reasoning

Yifeng Zhang, Ming Jiang, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15576-15585

Explainable visual question answering (VQA) models have been developed with neural modules and query-based knowledge incorporation to answer knowledge-requiring questions. Yet, most reasoning methods cannot effectively generate queries or incorporate external knowledge during the reasoning process, which may lead to suboptimal results. To bridge this research gap, we present Query and Attention Augmentation, a general approach that augments neural module networks to jointly reason about visual and external knowledge. To take both knowledge sources into account during reasoning, it parses the input question into a functional program with queries augmented through a novel reinforcement learning method, and jointly directs augmented attention to visual and external knowledge based on intermediate reasoning results. With extensive experiments on multiple VQA datasets, our method demonstrates significant performance, explainability, and generalizability over state-of-the-art models in answering questions requiring different extents of knowledge. Our source code is available at https://github.com/SuperJohnZhang/QAA.

****

Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, Candace Ross; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5238-5248

We present a novel task and dataset for evaluating the ability of vision and language models to conduct visio-linguistic compositional reasoning, which we call Winoground. Given two images and two captions, the goal is to match them correctly--but crucially, both captions contain a completely identical set of words, only in a different order. The dataset was carefully hand-curated by expert annotators and is labeled with a rich set of fine-grained tags to assist in analyzing

model performance. We probe a diverse range of state-of-the-art vision and langu age models and find that, surprisingly, none of them do much better than chance. Evidently, these models are not as skilled at visio-linguistic compositional re asoning as we might have hoped. We perform an extensive analysis to obtain insig hts into how future work might try to mitigate these models' shortcomings. We ai m for Winoground to serve as a useful evaluation set for advancing the state of the art and driving further progress in the field. The dataset is available at h ttps://huggingface.co/datasets/facebook/winoground.
****

RFNet: Unsupervised Network for Mutually Reinforcing Multi-Modal Image Registrat ion and Fusion

Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19679-1 9688

In this paper, we propose a novel method to realize multi-modal image registrati on and fusion in a mutually reinforcing framework, termed as RFNet. We handle th e registration in a coarse-to-fine fashion. For the first time, we exploit the f eedback of image fusion to promote the registration accuracy rather than treatin g them as two separate issues. The fine-registered results also improve the fusi on performance. Specifically, for image registration, we solve the bottlenecks o f defining registration metrics applicable for multi-modal images and facilitati ng the network convergence. The metrics are defined based on image translation a nd image fusion respectively in the coarse and fine stages. The convergence is f acilitated by the designed metrics and a deformable convolution-based network. F or image fusion, we focus on texture preservation, which not only increases the information amount and quality of fusion results but also improves the feedback of fusion results. The proposed method is evaluated on multi-modal images with l arge global parallaxes, images with local misalignments and aligned images to va lidate the performances of registration and fusion. The results in these cases d emonstrate the effectiveness of our method.
****

Progressive Attention on Multi-Level Dense Difference Maps for Generic Event Bou ndary Detection

Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, Limin Wang; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3 355-3364

Generic event boundary detection is an important yet challenging task in video u nderstanding, which aims at detecting the moments where humans naturally perceiv e event boundaries. The main challenge of this task is perceiving various tempor al variations of diverse event boundaries. To this end, this paper presents an e ffective and end-to-end learnable framework (DDM-Net). To tackle the diversity a nd complicated semantics of event boundaries, we make three notable improvements . First, we construct a feature bank to store multi-level features of space and time, prepared for difference calculation at multiple scales. Second, to allevia te inadequate temporal modeling of previous methods, we present dense difference maps (DDM) to comprehensively characterize the motion pattern. Finally, we expl oit progressive attention on multi-level DDM to jointly aggregate appearance and motion clues. As a result, DDM-Net respectively achieves a significant boost of 14% and 8% on Kinetics-GEBD and TAPOS benchmark, and outperforms the top-1 winn er solution of LOVEU Challenge@CVPR 2021 without bells and whistles. The state-o f-the-art result demonstrates the effectiveness of richer motion representation and more sophisticated aggregation, in handling the diversity of generic event b oundary detection. The code is made available at https://github.com/MCG-NJU/DDM.
****

Interactron: Embodied Adaptive Object Detection

Klemen Kotar, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2022, pp. 14860-14869

Over the years various methods have been proposed for the problem of object dete ction. Recently, we have witnessed great strides in this domain owing to the eme rgence of powerful deep neural networks. However, there are typically two main a

ssumptions common among these approaches. First, the model is trained on a fixed training set and is evaluated on a pre-recorded test set. Second, the model is kept frozen after the training phase, so no further updates are performed after the training is finished. These two assumptions limit the applicability of these methods to real-world settings. In this paper, we propose Interactron, a method for adaptive object detection in an interactive setting, where the goal is to perform object detection in images observed by an embodied agent navigating in different environments. Our idea is to continue training during inference and adapt the model at test time without any explicit supervision via interacting with the environment. Our adaptive object detection model provides a 11.8 point improvement in AP (and 19.1 points in AP50) over DETR, a recent, high-performance object detector. Moreover, we show that our object detection model adapts to environments with completely different appearance characteristics, and its performance is on par with a model trained with full supervision for those environments. We will release the code to help ease future research in this domain.
****

3D Scene Painting via Semantic Image Synthesis
Jaebong Jeong, Janghun Jo, Sunghyun Cho, Jaesik Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2262-2272
We propose a novel approach to 3D scene painting using a configurable 3D scene layout. Our approach takes a 3D scene with semantic class labels as input and trains a 3D scene painting network that synthesizes color values for the input 3D scene. We exploit an off-the-shelf 2D semantic image synthesis method to teach the 3D painting network without explicit color supervision. Experiments show that our approach produces images with geometrically correct structures and supports scene manipulation, such as the change of viewpoint, object poses, and painting style. Our approach provides rich controllability to synthesized images in the aspect of 3D geometry.
****

MeMOT: Multi-Object Tracking With Memory
Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8090-8100
We propose an online tracking algorithm that performs the object detection and data association under a common framework, capable of linking objects after a long time span. This is realized by preserving a large spatio-temporal memory to store the identity embeddings of the tracked objects, and by adaptively referencing and aggregating useful information from the memory as needed. Our model, called MeMOT, consists of three main modules that are all Transformer-based: 1) Hypothesis Generation that produce object proposals in the current video frame; 2) Memory Encoding that extracts the core information from the memory for each tracked object; and 3) Memory Decoding that solves the object detection and data association tasks simultaneously for multi-object tracking. When evaluated on widely adopted MOT benchmark datasets, MeMOT observes very competitive performance.
****

Revisiting Weakly Supervised Pre-Training of Visual Perception Models
Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, Laurens van der Maaten; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 804-814
Model pre-training is a cornerstone of modern visual recognition systems. Although fully supervised pre-training on datasets like ImageNet is still the de-facto standard, recent studies suggest that large-scale weakly supervised pre-training can outperform fully supervised approaches. This paper revisits weakly-supervised pre-training of models using hashtag supervision with modern versions of residual networks and the largest-ever dataset of images and corresponding hashtags. We study the performance of the resulting models in various transfer-learning settings including zero-shot transfer. We also compare our models with those obtained via large-scale self-supervised learning. We find our weakly-supervised mo

dels to be very competitive across all settings, and find they substantially out perform their self-supervised counterparts. We also include an investigation int o whether our models learned potentially troubling associations or stereotypes. Overall, our results provide a compelling argument for the use of weakly supervi sed learning in the development of visual recognition systems. Our models, Super vised Weakly through hashtAGs (SWAG), are available publicly.
****

Semi-Supervised Semantic Segmentation With Error Localization Network
Donghyeon Kwon, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 9957-9967
This paper studies semi-supervised learning of semantic segmentation, which assu mes that only a small portion of training images are labeled and the others rema in unlabeled. The unlabeled images are usually assigned pseudo labels to be used in training, which however often causes the risk of performance degradation due to the confirmation bias towards errors on the pseudo labels. We present a nove l method that resolves this chronic issue of pseudo labeling. At the heart of ou r method lies error localization network (ELN), an auxiliary module that takes a n image and its segmentation prediction as input and identifies pixels whose pse udo labels are likely to be wrong. ELN enables semi-supervised learning to be ro bust against inaccurate pseudo labels by disregarding label noises during traini ng and can be naturally integrated with self-training and contrastive learning. Moreover, we introduce a new learning strategy for ELN that simulates plausible and diverse segmentation errors during training of ELN to enhance its generaliza tion. Our method is evaluated on PASCAL VOC 2012 and Cityscapes, where it outper forms all existing methods in every evaluation setting.
****

Meta Convolutional Neural Networks for Single Domain Generalization
Chaoqun Wan, Xu Shen, Yonggang Zhang, Zhiheng Yin, Xinmei Tian, Feng Gao, Jianqi ang Huang, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 4682-4691
In single domain generalization, models trained with data from only one domain a re required to perform well on many unseen domains. In this paper, we propose a new model, termed meta convolutional neural network, to solve the single domain generalization problem in image recognition. The key idea is to decompose the co nvolutional features of images into meta features. Acting as "visual words", met a features are defined as universal and basic visual elements for image represen tations (like words for documents in language). Taking meta features as referenc e, we propose compositional operations to eliminate irrelevant features of local convolutional features by an addressing process and then to reformulate the con volutional feature maps as a composition of related meta features. In this way, images are universally coded without biased information from the unseen domain, which can be processed by following modules trained in the source domain. The co mpositional operations adopt a regression analysis technique to learn the meta f eatures in an online batch learning manner. Extensive experiments on multiple be nchmark datasets verify the superiority of the proposed model in improving singl e domain generalization ability.
****

Generalizing Gaze Estimation With Rotation Consistency
Yiwei Bao, Yunfei Liu, Haofei Wang, Feng Lu; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4207-4216
Recent advances of deep learning-based approaches have achieved remarkable perfo rmance on appearance-based gaze estimation. However, due to the shortage of targ et domain data and absence of target labels, generalizing gaze estimation algori thm to unseen environments is still challenging. In this paper, we discover the rotation-consistency property in gaze estimation and introduce the 'sub-label' f or unsupervised domain adaptation. Consequently, we propose the Rotation-enhance d Unsupervised Domain Adaptation (RUDA) for gaze estimation. First, we rotate th e original images with different angles for training. Then we conduct domain ada ptation under the constraint of rotation consistency. The target domain images a re assigned with sub-labels, derived from relative rotation angles rather than u

ntouchable real labels. With such sub-labels, we propose a novel distribution lo ss that facilitates the domain adaptation. We evaluate the RUDA framework on fou r cross-domain gaze estimation tasks. Experimental results demonstrate that it i mproves the performance over the baselines with gains ranging from 12.2% to 30.5 %. Our framework has the potential to be used in other computer vision tasks wit h physical constraints.
****

Anomaly Detection via Reverse Distillation From One-Class Embedding
Hanqiu Deng, Xingyu Li; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 9737-9746
Knowledge distillation (KD) achieves promising results on the challenging proble m of unsupervised anomaly detection (AD). The representation discrepancy of anom alies in the teacher-student (T-S) model provides essential evidence for AD. How ever, using similar or identical architectures to build the teacher and student models in previous studies hinders the diversity of anomalous representations. T o tackle this problem, we propose a novel T-S model consisting of a teacher enco der and a student decoder and introduce a simple yet effective "reverse distilla tion" paradigm accordingly. Instead of receiving raw images directly, the studen t network takes teacher model's one-class embedding as input and targets to rest ore the teacher's multi-scale representations. Inherently, knowledge distillatio n in this study starts from abstract, high-level presentations to low-level feat ures. In addition, we introduce a trainable one-class bottleneck embedding (OCBE) module in our T-S model. The obtained compact embedding effectively preserves essential information on normal patterns, but abandons anomaly perturbations. Ex tensive experimentation on AD and one-class novelty detection benchmarks shows t hat our method surpasses SOTA performance, demonstrating our proposed approach's effectiveness and generalizability.
****

Fine-Grained Object Classification via Self-Supervised Pose Alignment
Xuhui Yang, Yaowei Wang, Ke Chen, Yong Xu, Yonghong Tian; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 73 99-7408
Semantic patterns of fine-grained objects are determined by subtle appearance di fference of local parts, which thus inspires a number of part-based methods. How ever, due to uncontrollable object poses in images, distinctive details carried by local regions can be spatially distributed or even self-occluded, leading to a large variation on object representation. For discounting pose variations, thi s paper proposes to learn a novel graph based object representation to reveal a global configuration of local parts for self-supervised pose alignment across cl asses, which is employed as an auxiliary feature regularization on a deep repres entation learning network. Moreover, a coarse-to-fine supervision together with the proposed pose-insensitive constraint on shallow-to-deep sub-networks encoura ges discriminative features in a curriculum learning manner. We evaluate our met hod on three popular fine-grained object classification benchmarks, consistently achieving the state-of-the-art performance. Source codes are available at https ://github.com/yangxh11/P2P-Net.
****

Spatio-Temporal Gating-Adjacency GCN for Human Motion Prediction
Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, Shihong Xia; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6447-6456
Predicting future motion based on historical motion sequence is a fundamental pr oblem in computer vision, and it has wide applications in autonomous driving and robotics. Some recent works have shown that Graph Convolutional Networks(GCN) a re instrumental in modeling the relationship between different joints. However, considering the variants and diverse action types in human motion data, the cros s-dependency of the spatio-temporal relationships will be difficult to depict du e to the decoupled modeling strategy, which may also exacerbate the problem of i nsufficient generalization. Therefore, we propose the Spatio-Temporal Gating-Adj acency GCN(GAGCN) to learn the complex spatio-temporal dependencies over diverse

action types. Specifically, we adopt gating networks to enhance the generalizat
ion of GCN via the trainable adaptive adjacency matrix obtained by blending the
candidate spatio-temporal adjacency matrices. Moreover, GAGCN addresses the cros
s-dependency of space and time by balancing the weights of spatio-temporal model
ing and fusing the decoupled spatio-temporal features. Extensive experiments on
Human 3.6M, AMASS, and 3DPW demonstrate that GAGCN achieves state-of-the-art per
formance in both short-term and long-term predictions.
****

CellTypeGraph: A New Geometric Computer Vision Benchmark
Lorenzo Cerrone, Athul Vijayan, Tejasvinee Mody, Kay Schneitz, Fred A. Hamprecht
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 20897-20907
Classifying all cells in an organ is a relevant and difficult problem from plant
developmental biology. We here abstract the problem into a new benchmark for no
de classification in a geo-referenced graph. Solving it requires learning the sp
atial layout of the organ including symmetries. To allow the convenient testing
of new geometrical learning methods, the benchmark of Arabidopsis thaliana ovule
s is made available as a PyTorch data loader, along with a large number of preco
mputed features. Finally, we benchmark eight recent graph neural network archite
ctures, finding that DeeperGCN currently works best on this problem.
****

Clustering Plotted Data by Image Segmentation
Tarek Naous, Srinjay Sarkar, Abubakar Abid, James Zou; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21499
-21504
Clustering is a popular approach to detecting patterns in unlabeled data. Existi
ng clustering methods typically treat samples in a dataset as points in a metric
space and compute distances to group together similar points. In this paper, we
present a different way of clustering points in 2-dimensional space, inspired b
y how humans cluster data: by training neural networks to perform instance segme
ntation on plotted data. Our approach, Visual Clustering, has several advantages
over traditional clustering algorithms: it is much faster than most existing cl
ustering algorithms (making it suitable for very large datasets), it agrees stro
ngly with human intuition for clusters, and it is by default hyperparameter free
(although additional steps with hyperparameters can be introduced for more cont
rol of the algorithm). We describe the method and compare it to ten other cluste
ring methods on synthetic data to illustrate its advantages and disadvantages. W
e then demonstrate how our approach can be extended to higher-dimensional data a
nd illustrate its performance on real-world data. Our implementation of Visual C
lustering is publicly available as a python package that can be installed and us
ed on any dataset in a few lines of code. A demo on synthetic datasets is provid
ed.
****
****

Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding
Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, Jun Liu; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2022, pp. 19023-19034
Understanding animals' behaviors is significant for a wide range of applications
. However, existing animal behavior datasets have limitations in multiple aspect
s, including limited numbers of animal classes, data samples and provided tasks,
and also limited variations in environmental conditions and viewpoints. To addr
ess these limitations, we create a large and diverse dataset, Animal Kingdom, th
at provides multiple annotated tasks to enable a more thorough understanding of
natural animal behaviors. The wild animal footages used in our dataset record di
fferent times of the day in extensive range of environments containing variation
s in backgrounds, viewpoints, illumination and weather conditions. More specific
ally, our dataset contains 50 hours of annotated videos to localize relevant ani
mal behavior segments in long videos for the video grounding task, 30K video seq
uences for the fine-grained multi-label action recognition task, and 33K frames

for the pose estimation task, which correspond to a diverse range of animals with 850 species across 6 major animal classes. Such a challenging and comprehensive dataset shall be able to facilitate the community to develop, adapt, and evaluate various types of advanced methods for animal behavior analysis. Moreover, we propose a Collaborative Action Recognition (CARe) model that learns general and specific features for action recognition with unseen new animals. This method achieves promising performance in our experiments. Our dataset can be found at https://sutdcv.github.io/Animal-Kingdom.
****

Learning To Learn Across Diverse Data Biases in Deep Face Recognition
Chang Liu, Xiang Yu, Yi-Hsuan Tsai, Masoud Faraki, Ramin Moslemi, Manmohan Chandraker, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4072-4082
Convolutional Neural Networks have achieved remarkable success in face recognition, in part due to the abundant availability of data. However, the data used for training CNNs is often imbalanced. Prior works largely focus on the long-tailed nature of face datasets in data volume per identity, or focus on single bias variation. In this paper, we show that many bias variations such as ethnicity, head pose, occlusion and blur can jointly affect the accuracy significantly. We propose a sample level weighting approach termed Multi-variation Cosine Margin (MvCoM), to simultaneously consider the multiple variation factors, which orthogonally enhances the face recognition losses to incorporate the importance of training samples. Further, we leverage a learning to learn approach, guided by a held-out meta learning set and use an additive modeling to predict the MvCoM. Extensive experiments on challenging face recognition benchmarks demonstrate the advantages of our method in jointly handling imbalances due to multiple variations.
****

Back to Reality: Weakly-Supervised 3D Object Detection With Shape-Guided Label Enhancement
Xiuwei Xu, Yifan Wang, Yu Zheng, Yongming Rao, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8438-8447
In this paper, we propose a weakly-supervised approach for 3D object detection, which makes it possible to train a strong 3D detector with position-level annotations (i.e. annotations of object centers). In order to remedy the information loss from box annotations to centers, our method, namely Back to Reality (BR), makes use of synthetic 3D shapes to convert the weak labels into fully-annotated virtual scenes as stronger supervision, and in turn utilizes the perfect virtual labels to complement and refine the real labels. Specifically, we first assemble 3D shapes into physically reasonable virtual scenes according to the coarse scene layout extracted from position-level annotations. Then we go back to reality by applying a virtual-to-real domain adaptation method, which refine the weak labels and additionally supervise the training of detector with the virtual scenes. With less than 5% of the labeling labor, we achieve comparable detection performance with some popular fully-supervised approaches on the widely used ScanNet dataset.
****

Long-Tail Recognition via Compositional Knowledge Transfer
Sarah Parisot, Pedro M. Esperança, Steven McDonagh, Tamas J. Madarasz, Yongxin Yang, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6939-6948
In this work, we introduce a novel strategy for long-tail recognition that addresses the tail classes' few-shot problem via training-free knowledge transfer. Our objective is to transfer knowledge acquired from information-rich common classes to semantically similar, and yet data-hungry, rare classes in order to obtain stronger tail class representations. We leverage the fact that class prototypes and learned cosine classifiers provide two different, complementary representations of class cluster centres in feature space, and use an attention mechanism to select and recompose learned classifiers features from common classes to obtain higher quality rare class representations. Our knowledge transfer process is t

raining free, reducing overfitting risks, and can afford continual extension of classifiers to new classes. Experiments show that our approach can achieve significant performance boosts on rare classes while maintaining robust common class performance, outperforming directly comparable state-of-the-art models.
****

EI-CLIP: Entity-Aware Interventional Contrastive Learning for E-Commerce Cross-Modal Retrieval
Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, Xiaohui Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18051-18061
recommendation, and marketing services. Extensive efforts have been made to conquer the cross-modal retrieval problem in the general domain. When it comes to E-commerce, a common practice is to adopt the pretrained model and finetune on E-commerce data. Despite its simplicity, the performance is sub-optimal due to overlooking the uniqueness of E-commerce multimodal data. A few recent efforts have shown significant improvements over generic methods with customized designs for handling product images. Unfortunately, to the best of our knowledge, no existing method has addressed the unique challenges in the e-commerce language. This work studies the outstanding one, where it has a large collection of special meaning entities, e.g., "Dissel (brand)", "Top (category)", "relaxed (fit)" in the fashion clothing business. By formulating such out-of-distribution finetuning process in the Causal Inference paradigm, we view the erroneous semantics of these special entities as confounders to cause the retrieval failure. To rectify these semantics for aligning with e-commerce domain knowledge, we propose an intervention-based entity-aware contrastive learning framework with two modules, i.e., the Confounding Entity Selection Module and Entity-Aware Learning Module. Our method achieves competitive performance on the E-commerce benchmark Fashion-Gen. Particularly, in top-1 accuracy (R@1), we observe 10.3% and 10.5% relative improvements over the closest baseline in image-to-text and text-to-image retrievals, respectively.
****

Multi-Dimensional, Nuanced and Subjective - Measuring the Perception of Facial Expressions
De'Aira Bryant, Siqi Deng, Nashlie Sephus, Wei Xia, Pietro Perona; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20932-20941
Humans can perceive multiple expressions, each one with varying intensity, in the picture of a face. We propose a methodology for collecting and modeling multidimensional modulated expression annotations from human annotators. Our data reveals that the perception of some expressions can be quite different across observers; thus, our model is designed to represent ambiguity alongside intensity. An empirical exploration of how many dimensions are necessary to capture the perception of facial expression suggests six principal expression dimensions are sufficient. Using our method, we collected multidimensional modulated expression annotations for 1,000 images culled from the popular ExpW in-the-wild dataset. As a proof of principle of our improved measurement technique, we used these annotations to benchmark four public domain algorithms for automated facial expression prediction.
****

PyMiceTracking: An Open-Source Toolbox for Real-Time Behavioral Neuroscience Experiments
Richardson Menezes, Aron de Miranda, Helton Maia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21459-21465
The development of computational tools allows the advancement of research in behavioral neuroscience and elevates the limits of experiment design. Many behavioral experiments need to determine the animal's position from its tracking, which is crucial for real-time decision-making and further analysis of experimental data. Modern experimental designs usually generate the recording of a large amount of data, requiring the development of automatic computational tools and intelli

gent algorithms for timely data acquisition and processing. The proposed tool in this study initially operates with the acquisition of images. Then the animal tracking step begins with background subtraction, followed by the animal contour detection and morphological operations to remove noise in the detected shapes. Finally, in the final stage of the algorithm, the principal components analysis (PCA) is applied in the obtained shape, resulting in the animal's gaze direction.
****

Self-Taught Metric Learning Without Labels
Sungyeon Kim, Dongwon Kim, Minsu Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7431-7441
We present a novel self-taught framework for unsupervised metric learning, which alternates between predicting class-equivalence relations between data through a moving average of an embedding model and learning the model with the predicted relations as pseudo labels. At the heart of our framework lies an algorithm that investigates contexts of data on the embedding space to predict their class-equivalence relations as pseudo labels. The algorithm enables efficient end-to-end training since it demands no off-the-shelf module for pseudo labeling. Also, the class-equivalence relations provide rich supervisory signals for learning an embedding space. On standard benchmarks for metric learning, it clearly outperforms existing unsupervised learning methods and sometimes even beats supervised learning models using the same backbone network. It is also applied to semi-supervised metric learning as a way of exploiting additional unlabeled data, and achieves the state of the art by boosting performance of supervised learning substantially.
****

MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition
Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, Christoph Feichtenhofer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13587-13597
While today's video recognition systems parse snapshots or short clips accurately, they cannot connect the dots and reason across a longer range of time yet. Most existing video architectures can only process <5 seconds of a video without hitting the computation or memory bottlenecks. In this paper, we propose a new strategy to overcome this challenge. Instead of trying to process more frames at once like most existing methods, we propose to process videos in an online fashion and cache "memory" at each iteration. Through the memory, the model can reference prior context for long-term modeling, with only a marginal cost. Based on this idea, we build MeMViT, a Memory-augmented Multiscale Vision Transformer, that has a temporal support 30x longer than existing models with only 4.5 more compute; traditional methods need >3,000% more compute to do the same. On a wide range of settings, the increased temporal support enabled by MeMViT brings large gains in recognition accuracy consistently. MeMViT obtains state-of-the-art results on the AVA, EPIC-Kitchens-100 action classification, and action anticipation datasets.
****

Fine-Grained Temporal Contrastive Learning for Weakly-Supervised Temporal Action Localization
Junyu Gao, Mengyuan Chen, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19999-20009
We target at the task of weakly-supervised action localization (WSAL), where only video-level action labels are available during model training. Despite the recent progress, existing methods mainly embrace a localization-by-classification paradigm and overlook the fruitful fine-grained temporal distinctions between video sequences, thus suffering from severe ambiguity in classification learning and classification-to-localization adaption. This paper argues that learning by contextually comparing sequence-to-sequence distinctions offers an essential inductive bias in WSAL and helps identify coherent action instances. Specifically, under a differentiable dynamic programming formulation, two complementary contrastive objectives are designed, including Fine-grained Sequence Distance (FSD) cont

rasting and Longest Common Subsequence (LCS) contrasting, where the first one co nsiders the relations of various action/background proposals by using match, ins ert, and delete operators and the second one mines the longest common subsequenc es between two videos. Both contrasting modules can enhance each other and joint ly enjoy the merits of discriminative action-background separation and alleviate d task gap between classification and localization. Extensive experiments show t hat our method achieves state-of-the-art performance on two popular benchmarks. Our code is available at https://github.com/MengyuanChen21/CVPR2022-FTCL.
****

Embracing Single Stride 3D Object Detector With Sparse Transformer
Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8458-8468
In LiDAR-based 3D object detection for autonomous driving, the ratio of the obje ct size to input scene size is significantly smaller compared to 2D detection ca ses. Overlooking this difference, many 3D detectors directly follow the common p ractice of 2D detectors, which downsample the feature maps even after quantizing the point clouds. In this paper, we start by rethinking how such multi-stride s tereotype affects the LiDAR-based 3D object detectors. Our experiments point out that the downsampling operations bring few advantages, and lead to inevitable i nformation loss. To remedy this issue, we propose Single-stride Sparse Transform er (SST) to maintain the original resolution from the beginning to the end of th e network. Armed with transformers, our method addresses the problem of insuffic ient receptive field in single-stride architectures. It also cooperates well wit h the sparsity of point clouds and naturally avoids expensive computation. Event ually, our SST achieves state-of-the-art results on the large-scale Waymo Open D ataset. It is worth mentioning that our method can achieve exciting performance (83.8 LEVEL_1 AP on validation split) on small object (pedestrian) detection due to the characteristic of single stride. Our codes will be public soon.
****

Multidimensional Belief Quantification for Label-Efficient Meta-Learning
Deep Shankar Pandey, Qi Yu; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 14391-14400
Optimization-based meta-learning offers a promising direction for few-shot learn ing that is essential for many real-world computer vision applications. However, learning from few samples introduces uncertainty, and quantifying model confide nce for few-shot predictions is essential for many critical domains. Furthermore , few-shot tasks used in meta training are usually sampled randomly from a task distribution for an iterative model update, leading to high labeling costs and c omputational overhead in meta-training. We propose a novel uncertainty-aware tas k selection model for label efficient meta-learning. The proposed model formulat es a multidimensional belief measure, which can quantify the known uncertainty a nd lower bound the unknown uncertainty of any given task. Our theoretical result establishes an important relationship between the conflicting belief and the in correct belief. The theoretical result allows us to estimate the total uncertain ty of a task, which provides a principled criterion for task selection. A novel multi-query task formulation is further developed to improve both the computatio nal and labeling efficiency of meta-learning. Experiments conducted over multipl e real-world few-shot image classification tasks demonstrate the effectiveness o f the proposed model.
****

UTC: A Unified Transformer With Inter-Task Contrastive Learning for Visual Dialo g
Cheng Chen, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, Yudong Zhu, Xiaodo ng Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 18103-18112
Visual Dialog aims to answer multi-round, interactive questions based on the dia log history and image content. Existing methods either consider answer ranking a nd generating individually or only weakly capture the relation across the two ta sks implicitly by two separate models. The research on a universal framework tha

t jointly learns to rank and generate answers in a single model is seldom explor ed. In this paper, we propose a contrastive learning-based framework UTC to unif y and facilitate both discriminative and generative tasks in visual dialog with a single model. Specifically, considering the inherent limitation of the previou s learning paradigm, we devise two inter-task contrastive losses i.e., context c ontrastive loss and answer contrastive loss to make the discriminative and gener ative tasks mutually reinforce each other. These two complementary contrastive l osses exploit dialog context and target answer as anchor points to provide repre sentation learning signals from different perspectives. We evaluate our proposed UTC on the VisDial v1.0 dataset, where our method outperforms the state-of-the-art on both discriminative and generative tasks and surpasses previous state-of-the-art generative methods by more than 2 absolute points on Recall@1.
****

Relieving Long-Tailed Instance Segmentation via Pairwise Class Balance
Yin-Yin He, Peizhen Zhang, Xiu-Shen Wei, Xiangyu Zhang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 7000-7009
Long-tailed instance segmentation is a challenging task due to the extreme imbal ance of training samples among classes. It causes severe biases of the head clas ses (with majority samples) against the tailed ones. This renders "how to approp riately define and alleviate the bias" one of the most important issues. Prior w orks mainly use label distribution or mean score information to indicate a coars e-grained bias. In this paper, we explore to excavate the confusion matrix, whic h carries the fine-grained misclassification details, to relieve the pairwise bi ases, generalizing the coarse one. To this end, we propose a novel Pairwise Clas s Balance (PCB) method, built upon a confusion matrix which is updated during tr aining to accumulate the ongoing prediction preferences. PCB generates fightback soft labels for regularization during training. Besides, an iterative learning paradigm is developed to support a progressive and smooth regularization in such debiasing. PCB can be plugged and played to any existing methods as a complemen t. Experiments results on LVIS demonstrate that our method achieves state-of-the -art performance without bells and whistles. Superior results across various arc hitectures show the generalization ability. The code and trained models are avai lable at https://github.com/megvii-research/PCB.
****

Online Convolutional Re-Parameterization
Mu Hu, Junyi Feng, Jiashen Hua, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, Xia n-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 568-577
Structural re-parameterization has drawn increasing attention in various compute r vision tasks. It aims at improving the performance of deep models without intr oducing any inference-time cost. Though efficient during inference, such models rely heavily on the complicated training-time blocks to achieve high accuracy, l eading to large extra training cost. In this paper, we present online convolutio nal re-parameterization (OREPA), a two-stage pipeline, aiming to reduce the huge training overhead by squeezing the complex training-time block into a single co nvolution. To achieve this goal, we introduce a linear scaling layer for better optimizing the online blocks. Assisted with the reduced training cost, we also e xplore some more effective re-param components. Compared with the state-of-the-a rt re-param models, OREPA is able to save the training-time memory cost by about 70% and accelerate the training speed by around 2x. Meanwhile, equipped with OR EPA, the models outperform previous methods on ImageNet by up to +0.6%. We also conduct experiments on object detection and semantic segmentation and show consi stent improvements on the downstream tasks. Codes are available at https://githu b.com/JUGGHM/OREPA_CVPR2022.
****

Mimicking the Oracle: An Initial Phase Decorrelation Approach for Class Incremen tal Learning
Yujun Shi, Kuangqi Zhou, Jian Liang, Zihang Jiang, Jiashi Feng, Philip H.S. Torr , Song Bai, Vincent Y. F. Tan; Proceedings of the IEEE/CVF Conference on Compute

r Vision and Pattern Recognition (CVPR), 2022, pp. 16722-16731
Class Incremental Learning (CIL) aims at learning a classifier in a phase-by-pha se manner, in which only data of a subset of the classes are provided at each ph ase. Previous works mainly focus on mitigating forgetting in phases after the in itial one. However, we find that improving CIL at its initial phase is also a pr omising direction. Specifically, we experimentally show that directly encouragin g CIL Learner at the initial phase to output similar representations as the mode l jointly trained on all classes can greatly boost the CIL performance. Motivate d by this, we study the difference between a naively-trained initial-phase model and the oracle model. Specifically, since one major difference between these tw o models is the number of training classes, we investigate how such difference a ffects the model representations. We find that, with fewer training classes, the data representations of each class lie in a long and narrow region; with more t raining classes, the representations of each class scatter more uniformly. Inspi red by this observation, we propose Class-wise Decorrelation (CwD) that effectiv ely regularizes representations of each class to scatter more uniformly, thus mi micking the model jointly trained with all classes (i.e., the oracle model). Our CwD is simple to implement and easy to plug into existing methods. Extensive ex periments on various benchmark datasets show that CwD consistently and significa ntly improves the performance of existing state-of-the-art methods by around 1% to 3%. Code: https://github.com/Yujun-Shi/CwD.
****
RIDDLE: Lidar Data Compression With Range Image Deep Delta Encoding
Xuanyu Zhou, Charles R. Qi, Yin Zhou, Dragomir Anguelov; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 172 12-17221
Lidars are depth measuring sensors widely used in autonomous driving and augment ed reality. However, the large volume of data produced by lidars can lead to hig h costs in data storage and transmission. While lidar data can be represented as two interchangeable representations: 3D point clouds and range images, most pre vious work focus on compressing the generic 3D point clouds. In this work, we sh ow that directly compressing the range images can leverage the lidar scanning pa ttern, compared to compressing the unprojected point clouds. We propose a novel data-driven range image compression algorithm, named RIDDLE (Range Image Deep De Lta Encoding). At its core is a deep model that predicts the next pixel value in a raster scanning order, based on contextual laser shots from both the current and past scans (represented as a 4D point cloud of spherical coordinates and tim e). The deltas between predictions and original values can then be compressed by entropy encoding. Evaluated on the Waymo Open Dataset and KITTI, our method dem onstrates significant improvement in the compression rate (under the same distor tion) compared to widely used point cloud and range image compression algorithms as well as recent deep methods.
****
RelTransformer: A Transformer-Based Long-Tail Visual Relationship Recognition
Jun Chen, Aniket Agarwal, Sherif Abdelkarim, Deyao Zhu, Mohamed Elhoseiny; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 19507-19517
The visual relationship recognition (VRR) task aims at understanding the pairwis e visual relationships between interacting objects in an image. These relationsh ips typically have a long-tail distribution due to their compositional nature. T his problem gets more severe when the vocabulary becomes large, rendering this t ask very challenging. This paper shows that modeling an effective message-passin g flow through an attention mechanism can be critical to tackling the compositio nality and long-tail challenges in VRR. The method, called RelTransformer, repre sents each im- age as a fully-connected scene graph and restructures the whole s cene into the relation-triplet and global-scene contexts. It directly passes the message from each element in the relation-triplet and global-scene contexts to the target relation via self-attention. We also design a learnable memory to aug ment the long-tail relation representation learning. Through extensive experimen ts, we find that our model generalizes well on many VRR benchmarks. Our model ou

tperforms the best-performing models on two large-scale long-tail VRR benchmarks , VG8K-LT (+2.0% overall acc) and GQA-LT (+26.0% overall acc), both having a hig hly skewed distribution towards the tail. It also achieves strong results on the VG200 relation detection task. Our code is available at https://github.com/Visi on-CAIR/ RelTransformer.
****

HODEC: Towards Efficient High-Order DEcomposed Convolutional Neural Networks
Miao Yin, Yang Sui, Wanzhao Yang, Xiao Zang, Yu Gong, Bo Yuan; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 12299-12308
High-order decomposition is a widely used model compression approach towards com pact convolutional neural networks (CNNs). However, many of the existing solutio ns, though can efficiently reduce CNN model sizes, are very difficult to bring c onsiderable saving for computational costs, especially when the compression rati o is not huge, thereby causing the severe computation inefficiency problem. To o vercome this challenge, in this paper we propose efficient High-Order DEcomposed Convolution (HODEC). By performing systematic explorations on the underlying re ason and mitigation strategy for the computation inefficiency, we develop a new decomposition and computation-efficient execution scheme, enabling simultaneous reductions in computational and storage costs. To demonstrate the effectiveness of HODEC, we perform empirical evaluations for various CNN models on different d atasets. HODEC shows consistently outstanding compression and acceleration perfo rmance. For ResNet-56 on CIFAR-10 dataset, HODEC brings 67% fewer model paramete rs and 62% fewer FLOPs with 1.17% accuracy increase than the baseline. For ResNe t-50 on ImageNet dataset, HODEC achieves 63% FLOPs reduction with 0.31% accuracy increase than the uncompressed model.
****

RigidFlow: Self-Supervised Scene Flow Learning on Point Clouds by Local Rigidity Prior
Ruibo Li, Chi Zhang, Guosheng Lin, Zhe Wang, Chunhua Shen; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1 6959-16968
In this work, we focus on scene flow learning on point clouds in a self-supervis ed manner. A real-world scene can be well modeled as a collection of rigidly mov ing parts, therefore its scene flow can be represented as a combination of rigid motion of each part. Inspired by this observation, we propose to generate pseud o scene flow for self-supervised learning based on piecewise rigid motion estima tion, in which the source point cloud is decomposed into a set of local regions and each region is treated as rigid. By rigidly aligning each region with its po tential counterpart in the target point cloud, we obtain a region-specific rigid transformation to represent the flow, which together constitutes the pseudo sce ne flow labels of the entire scene to enable network training. Compared with mos t existing approaches relying on point-wise similarities for point matching, our method explicitly enforces region-wise rigid alignments, yielding locally rigid pseudo scene flow labels. We demonstrate the effectiveness of our self-supervis ed learning method on FlyingThings3D and KITTI datasets. Comprehensive experimen ts show that our method achieves new state-of-the-art performance in self-superv ised scene flow learning, without any ground truth scene flow for supervision, e ven outperforming some supervised counterparts.
****

Smooth Maximum Unit: Smooth Activation Function for Deep Networks Using Smoothin g Maximum Technique
Koushik Biswas, Sandeep Kumar, Shilpak Banerjee, Ashish Kumar Pandey; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 794-803
Deep learning researchers have a keen interest in proposing new novel activation functions that can boost neural network performance. A good choice of activatio n function can have a significant effect on improving network performance and tr aining dynamics. Rectified Linear Unit (ReLU) is a popular hand-designed activat ion function and is the most common choice in the deep learning community due to

its simplicity though ReLU has some drawbacks. In this paper, we have proposed two new novel activation functions based on approximation of the maximum function, and we call these functions Smooth Maximum Unit (SMU and SMU-1). We show that SMU and SMU-1 can smoothly approximate ReLU, Leaky ReLU, or more general Maxout family, and GELU is a particular case of SMU. Replacing ReLU by SMU, Top-1 classification accuracy improves by 6.22%, 3.39%, 3.51%, and 3.08% on the CIFAR100 dataset with ShuffleNet V2, PreActResNet-50, ResNet-50, and SeNet-50 models respectively. Also, our experimental evaluation shows that SMU and SMU-1 improve network performance in a variety of deep learning tasks like image classification, object detection, semantic segmentation, and machine translation compared to widely used activation functions.

****

Learning Invisible Markers for Hidden Codes in Offline-to-Online Photography
Jun Jia, Zhongpai Gao, Dandan Zhu, Xiongkuo Min, Guangtao Zhai, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2273-2282

QR (quick response) codes are widely used as an offline-to-online channel to convey information (e.g., links) from publicity materials (e.g., display and print) to mobile devices. However, QR Codes are not favorable for taking up valuable space of publicity materials. Recent works propose invisible codes/hyperlinks that can convey hidden information from offline to online. However, they require markers to locate invisible codes, which fails the purpose of invisible codes to be visible because of the markers. This paper proposes a novel invisible information hiding architecture for display/print-camera scenarios, consisting of hiding, locating, correcting, and recovery, where invisible markers are learned to make hidden codes truly invisible. We hide information in a sub-image rather than the entire image and include a localization module in the end-to-end framework. To achieve both high visual quality and high recovering robustness, an effective multi-stage training strategy is proposed. The experimental results show that the proposed method outperforms the state-of-the-art information hiding methods in both visual quality and robustness. In addition, the automatic localization of hidden codes significantly reduces the time of manually correcting geometric distortions for photos, which is a revolutionary innovation for information hiding in mobile applications.

****

Personalized Image Aesthetics Assessment With Rich Attributes
Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, Yandong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19861-19869

Personalized image aesthetics assessment (PIAA) is challenging due to its highly subjective nature. People's aesthetic tastes depend on diversified factors, including image characteristics and subject characters. The existing PIAA databases are limited in terms of annotation diversity, especially the subject aspect, which can no longer meet the increasing demands of PIAA research. To solve the dilemma, we conduct so far, the most comprehensive subjective study of personalized image aesthetics and introduce a new Personalized image Aesthetics database with Rich Attributes (PARA), which consists of 31,220 images with annotations by 438 subjects. PARA features wealthy annotations, including 9 image-oriented objective attributes and 4 human-oriented subjective attributes. In addition, desensitized subject information, such as personality traits, is also provided to support study of PIAA and user portraits. A comprehensive analysis of the annotation data is provided and statistic study indicates that the aesthetic preferences can be mirrored by proposed subjective attributes. We also propose a conditional PIAA model by utilizing subject information as conditional prior. Experimental results indicate that the conditional PIAA model can outperform the control group, which is also the first attempt to demonstrate how image aesthetics and subject characters interact to produce the intricate personalized tastes on image aesthetics. We believe the database and the associated analysis would be useful for conducting next-generation PIAA study. The project page of PARA can be found at https://cv-datasets.institutecv.com/#/data-sets.

****

Task2Sim: Towards Effective Pre-Training and Transfer From Synthetic Data
Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu (Richard) Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, Rogerio S. Feris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9194-9204

Pre-training models on Imagenet or other massive datasets of real images has led to major advances in computer vision, albeit accompanied with shortcomings related to curation cost, privacy, usage rights, and ethical issues. In this paper, for the first time, we study the transferability of pre-trained models based on synthetic data generated by graphics simulators to downstream tasks from very different domains. In using such synthetic data for pre-training, we find that downstream performance on different tasks are favored by different configurations of simulation parameters (e.g. lighting, object pose, backgrounds, etc.), and that there is no one-size-fits-all solution. It is thus better to tailor synthetic pre-training data to a specific downstream task, for best performance. We introduce Task2Sim, a unified model mapping downstream task representations to optimal simulation parameters to generate synthetic pre-training data for them. Task2Sim learns this mapping by training to find the set of best parameters on a set of "seen" tasks. Once trained, it can then be used to predict best simulation parameters for novel "unseen" tasks in one shot, without requiring additional training. Given a budget in number of images per class, our extensive experiments with 20 diverse downstream tasks show Task2Sim's task-adaptive pre-training data results in significantly better downstream performance than non-adaptively choosing simulation parameters on both seen and unseen tasks. It is even competitive with pre-training on real images from Imagenet.

****

Part-Based Pseudo Label Refinement for Unsupervised Person Re-Identification
Yoonki Cho, Woo Jae Kim, Seunghoon Hong, Sung-Eui Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7308-7318

Unsupervised person re-identification (re-ID) aims at learning discriminative representations for person retrieval from unlabeled data. Recent techniques accomplish this task by using pseudo-labels, but these labels are inherently noisy and deteriorate the accuracy. To overcome this problem, several pseudo-label refinement methods have been proposed, but they neglect the fine-grained local context essential for person re-ID. In this paper, we propose a novel Part-based Pseudo Label Refinement (PPLR) framework that reduces the label noise by employing the complementary relationship between global and part features. Specifically, we design a cross agreement score as the similarity of k-nearest neighbors between feature spaces to exploit the reliable complementary relationship. Based on the cross agreement, we refine pseudo-labels of global features by ensembling the predictions of part features, which collectively alleviate the noise in global feature clustering. We further refine pseudo-labels of part features by applying label smoothing according to the suitability of given labels for each part. Thanks to the reliable complementary information provided by the cross agreement score, our PPLR effectively reduces the influence of noisy labels and learns discriminative representations with rich local contexts. Extensive experimental results on Market-1501 and MSMT17 demonstrate the effectiveness of the proposed method over the state-of-the-art performance. The code is available at https://github.com/yoonkicho/PPLR.

****

Bridging the Gap Between Learning in Discrete and Continuous Environments for Vision-and-Language Navigation
Yicong Hong, Zun Wang, Qi Wu, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15439-15449

Most existing works in vision-and-language navigation (VLN) focus on either discrete or continuous environments, training agents that cannot generalize across the two. Although learning to navigate in continuous spaces is closer to the real-world, training such an agent is significantly more difficult than training an

agent in discrete spaces. However, recent advances in discrete VLN are challengi ng to translate to continuous VLN due to the domain gap. The fundamental differe nce between the two setups is that discrete navigation assumes prior knowledge o f the connectivity graph of the environment, so that the agent can effectively t ransfer the problem of navigation with low-level controls to jumping from node t o node with high-level actions by grounding to an image of a navigable direction . To bridge the discrete-to-continuous gap, we propose a predictor to generate a set of candidate waypoints during navigation, so that agents designed with high -level actions can be transferred to and trained in continuous environments. We refine the connectivity graph of Matterport3D to fit the continuous Habitat-Matt erport3D, and train the waypoints predictor with the refined graphs to produce a ccessible waypoints at each time step. Moreover, we demonstrate that the predict ed waypoints can be augmented during training to diversify the views and paths, and therefore enhance agent's generalization ability. Through extensive experime nts we show that agents navigating in continuous environments with predicted way points perform significantly better than agents using low-level actions, which r educes the absolute discrete-to-continuous gap by 11.76% Success Weighted by Pat h Length (SPL) for the Cross-Modal Matching Agent and 18.24% SPL for the Recurre nt VLN-BERT. Our agents, trained with a simple imitation learning objective, out perform previous methods by a large margin, achieving new state-of-the-art resul ts on the testing environments of the R2R-CE and the RxR-CE datasets.

****

HDNet: High-Resolution Dual-Domain Learning for Spectral Compressive Imaging

Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Tim ofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 17542-17551

The rapid development of deep learning provides a better solution for the end-to -end reconstruction of hyperspectral image (HSI). However, existing learning-bas ed methods have two major defects. Firstly, networks with self-attention usually sacrifice internal resolution to balance model performance against complexity, losing fine-grained high-resolution (HR) features. Secondly, even if the optimiz ation focusing on spatial-spectral domain learning (SDL) converges to the ideal solution, there is still a significant visual difference between the reconstruct ed HSI and the truth. So we propose a high-resolution dual-domain learning netwo rk (HDNet) for HSI reconstruction. On the one hand, the proposed HR spatial-spec tral attention module with its efficient feature fusion provides continuous and fine pixel-level features. On the other hand, frequency domain learning (FDL) is introduced for HSI reconstruction to narrow the frequency domain discrepancy. D ynamic FDL supervision forces the model to reconstruct fine-grained frequencies and compensate for excessive smoothing and distortion caused by pixel-level loss es. The HR pixel-level attention and frequency-level refinement in our HDNet mut ually promote HSI perceptual quality. Extensive quantitative and qualitative exp eriments show that our method achieves SOTA performance on simulated and real HS I datasets. https://github.com/Huxiaowan/HDNet

****

OW-DETR: Open-World Detection Transformer

Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, Muba rak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9235-9244

Open-world object detection (OWOD) is a challenging computer vision problem, whe re the task is to detect a known set of object categories while simultaneously i dentifying unknown objects. Additionally, the model must incrementally learn new classes that become known in the next training episodes. Distinct from standard object detection, the OWOD setting poses significant challenges for generating quality candidate proposals on potentially unknown objects, separating the unkno wn objects from the background and detecting diverse unknown objects. Here, we i ntroduce a novel end-to-end transformer-based framework, OW-DETR, for open-world object detection. The proposed OW-DETR comprises three dedicated components nam ely, attention-driven pseudo-labeling, novelty classification and objectness sco ring to explicitly address the aforementioned OWOD challenges. Our OW-DETR expli

citly encodes multi-scale contextual information, possesses less inductive bias, enables knowledge transfer from known classes to the unknown class and can better discriminate between unknown objects and background. Comprehensive experiments are performed on two benchmarks: MS-COCO and PASCAL VOC. The extensive ablations reveal the merits of our proposed contributions. Further, our model outperforms the recently introduced OWOD approach, ORE, with absolute gains ranging from 1.8% to 3.3% in terms of unknown recall on MS-COCO. In the case of incremental object detection, OW-DETR outperforms the state-of-the-art for all settings on PASCAL VOC. Our code is available at https://github.com/akshitac8/OW-DETR.

****

Learning Deep Implicit Functions for 3D Shapes With Dynamic Code Clouds

Tianyang Li, Xin Wen, Yu-Shen Liu, Hua Su, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12840-12850

Deep Implicit Function (DIF) has gained popularity as an efficient 3D shape representation. To capture geometry details, current methods usually learn DIF using local latent codes, which discretize the space into a regular 3D grid (or octree) and store local codes in grid points (or octree nodes). Given a query point, the local feature is computed by interpolating its neighboring local codes with their positions. However, the local codes are constrained at discrete and regular positions like grid points, which makes the code positions difficult to be optimized and limits their representation ability. To solve this problem, we propose to learn DIF with Dynamic Code Cloud, named DCC-DIF. Our method explicitly associates local codes with learnable position vectors, and the position vectors are continuous and can be dynamically optimized, which improves the representation ability. In addition, we propose a novel code position loss to optimize the code positions, which heuristically guides more local codes to be distributed around complex geometric details. In contrast to previous methods, our DCC-DIF represents 3D shapes more efficiently with a small amount of local codes, and improves the reconstruction quality. Experiments demonstrate that DCC-DIF achieves better performance over previous methods. Code and data are available at https://github.com/lity20/DCCDIF.

****

Reversible Vision Transformers

Karttikeya Mangalam, Haoqi Fan, Yanghao Li, Chao-Yuan Wu, Bo Xiong, Christoph Feichtenhofer, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10830-10840

We present Reversible Vision Transformers, a memory efficient architecture design for visual recognition. By decoupling the GPU memory footprint from the depth of the model, Reversible Vision Transformers enable memory efficient scaling of transformer architectures. We adapt two popular models, namely Vision Transformer and Multi-scale Vision Transformers, to reversible variants and benchmark extensively across both model sizes and tasks of image classification, object detection and video classification. Reversible Vision Transformers achieve a reduced memory footprint of up to 15.5x at identical model complexity, parameters and accuracy, demonstrating the promise of reversible vision transformers as an efficient backbone for resource limited training regimes. Finally, we find that the additional computational burden of recomputing activations is more than overcome for deeper models, where throughput can increase up to 3.9x over their non-reversible counterparts. Code and models are available at https://github.com/facebookresearch/mvit.

****

Amodal Panoptic Segmentation

Rohit Mohan, Abhinav Valada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21023-21032

Humans have the remarkable ability to perceive objects as a whole, even when parts of them are occluded. This ability of amodal perception forms the basis of our perceptual and cognitive understanding of our world. To enable robots to reason with this capability, we formulate and propose a novel task that we name amodal panoptic segmentation. The goal of this task is to simultaneously predict the

pixel-wise semantic segmentation labels of the visible regions of stuff classes and the instance segmentation labels of both the visible and occluded regions of thing classes. To facilitate research on this new task, we extend two established benchmark datasets with pixel-level amodal panoptic segmentation labels that we make publicly available as KITTI-360-APS and BDD100K-APS. We present several strong baselines, along with the amodal panoptic quality (APQ) and amodal parsing coverage (APC) metrics to quantify the performance in an interpretable manner. Furthermore, we propose the novel amodal panoptic segmentation network (APSNet), as a first step towards addressing this task by explicitly modeling the complex relationships between the occluders and occludes. Extensive experimental evaluations demonstrate that APSNet achieves state-of-the-art performance on both benchmarks and more importantly exemplifies the utility of amodal recognition. The datasets are available at http://amodal-panoptic.cs.uni-freiburg.de

****

Gravitationally Lensed Black Hole Emission Tomography

Aviad Levis, Pratul P. Srinivasan, Andrew A. Chael, Ren Ng, Katherine L. Bouman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19841-19850

Measurements from the Event Horizon Telescope enabled the visualization of light emission around a black hole for the first time. So far, these measurements have been used to recover a 2D image under the assumption that the emission field is static over the period of acquisition. In this work, we propose BH-NeRF, a novel tomography approach that leverages gravitational lensing to recover the continuous 3D emission field near a black hole. Compared to other 3D reconstruction or tomography settings, this task poses two significant challenges: first, rays near black holes follow curved paths dictated by general relativity, and second, we only observe measurements from a single viewpoint. Our method captures the unknown emission field using a continuous volumetric function parameterized by a coordinate-based neural network, and uses knowledge of Keplerian orbital dynamics to establish correspondence between 3D points over time. Together, these enable BH-NeRF to recover accurate 3D emission fields, even in challenging situations with sparse measurements and uncertain orbital dynamics. This work takes the first steps in showing how future measurements from the Event Horizon Telescope could be used to recover evolving 3D emission around the supermassive black hole in our Galactic center.

****

3D-Aware Image Synthesis via Learning Structural and Textural Representations

Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18430-18439

Making generative models 3D-aware bridges the 2D image space and the 3D physical world yet remains challenging. Recent attempts equip a Generative Adversarial Network (GAN) with a Neural Radiance Field (NeRF), which maps 3D coordinates to pixel values, as a 3D prior. However, the implicit function in NeRF has a very local receptive field, making the generator hard to become aware of the global structure. Meanwhile, NeRF is built on volume rendering which can be too costly to produce high-resolution results, increasing the optimization difficulty. To alleviate these two problems, we propose a novel framework, termed as VolumeGAN, for high-fidelity 3D-aware image synthesis, through explicitly learning a structural representation and a textural representation. We first learn a feature volume to represent the underlying structure, which is then converted to a feature field using a NeRF-like model. The feature field is further accumulated into a 2D feature map as the textural representation, followed by a neural renderer for appearance synthesis. Such a design enables independent control of the shape and the appearance. Extensive experiments on a wide range of datasets confirm that, our approach achieves sufficiently higher image quality and better 3D control than the previous methods..

****

Text-to-Image Synthesis Based on Object-Guided Joint-Decoding Transformer

Fuxiang Wu, Liu Liu, Fusheng Hao, Fengxiang He, Jun Cheng; Proceedings of the IE

EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18113-18122
Object-guided text-to-image synthesis aims to generate images from natural language descriptions built by two-step frameworks, i.e., the model generates the layout and then synthesizes images from the layout and captions. However, such frameworks have two issues: 1) complex structure, since generating language-related layout is not a trivial task; 2) error propagation, because the inappropriate layout will mislead the image synthesis and is hard to be revised. In this paper, we propose an object-guided joint-decoding module to simultaneously generate the image and the corresponding layout. Specially, we present the joint-decoding transformer to model the joint probability on images tokens and the corresponding layouts tokens, where layout tokens provide additional observed data to model the complex scene better. Then, we describe a novel Layout-VQGAN for layout encoding and decoding to provide more information about the complex scene. After that, we present the detail-enhanced module to enrich the language-related details based on two facts: 1) visual details could be omitted in the compression of VQGANs; 2) the joint-decoding transformer would not have sufficient generating capacity. The experiments show that our approach is competitive with previous object-centered models and can generate diverse and high-quality objects under the given layouts.
****

Correlation Verification for Image Retrieval
Seongwon Lee, Hongje Seong, Suhyeon Lee, Euntai Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5374-5384
Geometric verification is considered a de facto solution for the re-ranking task in image retrieval. In this study, we propose a novel image retrieval re-ranking network named Correlation Verification Networks (CVNet). Our proposed network, comprising deeply stacked 4D convolutional layers, gradually compresses dense feature correlation into image similarity while learning diverse geometric matching patterns from various image pairs. To enable cross-scale matching, it builds feature pyramids and constructs cross-scale feature correlations within a single inference, replacing costly multi-scale inferences. In addition, we use curriculum learning with the hard negative mining and Hide-and-Seek strategy to handle hard samples without losing generality. Our proposed re-ranking network shows state-of-the-art performance on several retrieval benchmarks with a significant margin (+12.6% in mAP on ROxford-Hard+1M set) over state-of-the-art methods. The source code and models are available online: https://github.com/sungonce/CVNet.
****

Unsupervised Vision-and-Language Pre-Training via Retrieval-Based Multi-Granular Alignment
Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, Ning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16485-16494
Vision-and-Language (V+L) pre-training models have achieved tremendous success in recent years on various multi-modal benchmarks. However, the majority of existing models require pre-training on a large set of parallel image-text data, which is costly to collect, compared to image-only or text-only data. In this paper, we propose unsupervised Vision-and-Language pre-training (UVLP) to learn the cross-modal representation from non-parallel image and text datasets. We found two key factors that lead to good unsupervised V+L pre-training without parallel data: (i) joint image-and-text input (ii) overall image-text alignment (even for non-parallel data). Accordingly, we propose a novel unsupervised V+L pre-training curriculum for non-parallel texts and images. We first construct a weakly aligned image-text corpus via a retrieval-based approach, then apply a set of multi-granular alignment pre-training tasks, including region-to-tag, region-to-phrase, and image-to-sentence alignment, to bridge the gap between the two modalities. A comprehensive ablation study shows each granularity is helpful to learn a stronger pre-trained model. We adapt our pre-trained model to a set of V+L downstream tasks, including VQA, NLVR2, Visual Entailment, and RefCOCO+. Our model achiev

es the state-of-art performance in all these tasks under the unsupervised settin
g.
****

Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-Robus
t Makeup Transfer

Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, Lib
ing Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2022, pp. 15014-15023

While deep face recognition (FR) systems have shown amazing performance in ident
ification and verification, they also arouse privacy concerns for their excessiv
e surveillance on users, especially for public face images widely spread on soci
al networks. Recently, some studies adopt adversarial examples to protect photos
 from being identified by unauthorized face recognition systems. However, existi
ng methods of generating adversarial face images suffer from many limitations, s
uch as awkward visual, white-box setting, weak transferability, making them diff
icult to be applied to protect face privacy in reality. In this paper, we propos
e adversarial makeup transfer GAN (AMT-GAN), a novel face protection method aimi
ng at constructing adversarial face images that preserve stronger black-box tran
sferability and better visual quality simultaneously. AMT-GAN leverages generati
ve adversarial networks (GAN) to synthesize adversarial face images with makeup
 transferred from reference images. In particular, we introduce a new regularizat
ion module along with a joint training strategy to reconcile the conflicts betwe
en the adversarial noises and the cycle consistence loss in makeup transfer, ach
ieving a desirable balance between the attack strength and visual changes. Exten
sive experiments verify that compared with state of the arts, AMT-GAN can not on
ly preserve a comfortable visual quality, but also achieve a higher attack succe
ss rate over commercial FR APIs, including Face++, Aliyun, and Microsoft.
****

PONI: Potential Functions for ObjectGoal Navigation With Interaction-Free Learni
ng

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Mal
ik, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2022, pp. 18890-18900

State-of-the-art approaches to ObjectGoal navigation (ObjectNav) rely on reinfor
cement learning and typically require significant computational resources and ti
me for learning. We propose Potential functions for ObjectGoal Navigation with I
nteraction-free learning (PONI), a modular approach that disentangles the skills
 of 'where to look?' for an object and 'how to navigate to (x, y)?'. Our key ins
ight is that 'where to look?' can be treated purely as a perception problem, and
 learned without environment interactions. To address this, we propose a network
 that predicts two complementary potential functions conditioned on a semantic m
ap and uses them to decide where to look for an unseen object. We train the pote
ntial function network using supervised learning on a passive dataset of top-dow
n semantic maps, and integrate it into a modular framework to perform ObjectNav.
 Experiments on Gibson and Matterport3D demonstrate that our method achieves the
 state-of-the-art for ObjectNav while incurring up to 1,600x less computational
cost for training. Code and pre-trained models are available.
****

Noise Is Also Useful: Negative Correlation-Steered Latent Contrastive Learning

Jiexi Yan, Lei Luo, Chenghao Xu, Cheng Deng, Heng Huang; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 31-
40

How to effectively handle label noise has been one of the most practical but cha
llenging tasks in Deep Neural Networks (DNNs). Recent popular methods for traini
ng DNNs with noisy labels mainly focus on directly filtering out samples with lo
w confidence or repeatedly mining valuable information from low-confident sample
s. %to further modify DNNs. However, they cannot guarantee the robust generaliza
tion of models due to the ignorance of useful information hidden in noisy data.
To address this issue, we propose a new effective method named as LaCoL (Latent
Contrastive Learning) to leverage the negative correlations from the noisy data.

Specifically, in label space, we exploit the weakly-augmented data to filter samples and adopt classification loss on strong augmentations of the selected sample set, which can preserve the training diversity. While in metric space, we utilize weakly-supervised contrastive learning to excavate these negative correlations hidden in noisy data. Moreover, a cross-space similarity consistency regularization is provided to constrain the gap between label space and metric space. Extensive experiments have validated the superiority of our approach over existing state-of-the-art methods.

****

Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation

Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, Xiang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11006-11016

Multi-frame human pose estimation has long been a compelling and fundamental problem in computer vision. This task is challenging due to fast motion and pose occlusion that frequently occur in videos. State-of-the-art methods strive to incorporate additional visual evidences from neighboring frames (supporting frames) to facilitate the pose estimation of the current frame (key frame). One aspect that has been obviated so far, is the fact that current methods directly aggregate unaligned contexts across frames. The spatial-misalignment between pose features of the current frame and neighboring frames might lead to unsatisfactory results. More importantly, existing approaches build upon the straightforward pose estimation loss, which unfortunately cannot constrain the network to fully leverage useful information from neighboring frames. To tackle these problems, we present a novel hierarchical alignment framework, which leverages coarse-to-fine deformations to progressively update a neighboring frame to align with the current frame at the feature level. We further propose to explicitly supervise the knowledge extraction from neighboring frames, guaranteeing that useful complementary cues are extracted. To achieve this goal, we theoretically analyzed the mutual information between the frames and arrived at a loss that maximizes the taskrelevant mutual information. These allow us to rank No.1 in the Multi-frame Person Pose Estimation Challenge on benchmark dataset PoseTrack2017, and obtain state-of-the-art performance on benchmarks Sub-JHMDB and PoseTrack2018. Our code is released at https://github.com/Pose-Group/FAMI-Pose, hoping that it will be useful to the community.

****

Spatially-Adaptive Multilayer Selection for GAN Inversion and Editing

Gaurav Parmar, Yijun Li, Jingwan Lu, Richard Zhang, Jun-Yan Zhu, Krishna Kumar Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11399-11409

Existing GAN inversion and editing methods work well for aligned objects with a clean background, such as portraits and animal faces, but often struggle for more difficult categories with complex scene layouts and object occlusions, such as cars, animals, and outdoor images. We propose a new method to invert and edit such complex images in the latent space of GANs, such as StyleGAN2. Our key idea is to explore inversion with a collection of layers, spatially adapting the inversion process to the difficulty of the image. We learn to predict the "invertibility" of different image segments and project each segment into a latent layer. Easier regions can be inverted into an earlier layer in the generator's latent space, while more challenging regions can be inverted into a later feature space. Experiments show that our method obtains better inversion results compared to the recent approaches on complex categories, while maintaining downstream editability. Please refer to our project page at gauravparmar.com/sam_ inversion.

****

Self-Supervised Transformers for Unsupervised Object Discovery Using Normalized Cut

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, Dominique Vaufreydaz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14543-14553

Transformers trained with self-supervision using self-distillation loss (DINO) have been shown to produce attention maps that highlight salient foreground objects. In this paper, we show a graph-based method that uses the self-supervised transformer features to discover an object from an image. Visual tokens are viewed as nodes in a weighted graph with edges representing a connectivity score based on the similarity of tokens. Foreground objects can then be segmented using a normalized graph-cut to group self-similar regions. We solve the graph-cut problem using spectral clustering with generalized eigen-decomposition and show that the second smallest eigenvector provides a cutting solution since its absolute value indicates the likelihood that a token belongs to a foreground object. Despite its simplicity, this approach significantly boosts the performance of unsupervised object discovery: we improve over the recent state-of-the-art LOST by a margin of 6.9%, 8.1%, and 8.1% respectively on the VOC07, VOC12, and COCO20K. The performance can be further improved by adding a second stage class-agnostic detector (CAD). Our proposed method can be easily extended to unsupervised saliency detection and weakly supervised object detection. For unsupervised saliency detection, we improve IoU for 4.9%, 5.2%, 12.9% on ECSSD, DUTS, DUTOMRON respectively compared to state-of-the-art. For weakly supervised object detection, we achieve competitive performance on CUB and ImageNet. Our code is available at: https://www.m-psi.fr/Papers/TokenCut2022/

****

Exploring Structure-Aware Transformer Over Interaction Proposals for Human-Object Interaction Detection

Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, Chang-Wen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19548-19557

Recent high-performing Human-Object Interaction (HOI) detection techniques have been highly influenced by Transformer-based object detector (i.e., DETR). Nevertheless, most of them directly map parametric interaction queries into a set of HOI predictions through vanilla Transformer in a one-stage manner. This leaves rich inter- or intra-interaction structure under-exploited. In this work, we design a novel Transformer-style HOI detector, i.e., Structure-aware Transformer over Interaction Proposals (STIP), for HOI detection. Such design decomposes the process of HOI set prediction into two subsequent phases, i.e., an interaction proposal generation is first performed, and then followed by transforming the non-parametric interaction proposals into HOI predictions via a structure-aware Transformer. The structure-aware Transformer upgrades vanilla Transformer by encoding additionally the holistically semantic structure among interaction proposals as well as the locally spatial structure of human/object within each interaction proposal, so as to strengthen HOI predictions. Extensive experiments conducted on V-COCO and HICO-DET benchmarks have demonstrated the effectiveness of STIP, and superior results are reported when comparing with the state-of-the-art HOI detectors. Source code is available at https://github.com/zyong812/STIP.

****

Towards Robust Adaptive Object Detection Under Noisy Annotations

Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, Yixuan Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14207-14216

Domain Adaptive Object Detection (DAOD) models a joint distribution of images and labels from an annotated source domain and learns a domain-invariant transformation to estimate the target labels with the given target domain images. Existing methods assume that the source domain labels are completely clean, yet large-scale datasets often contain error-prone annotations due to instance ambiguity, which may lead to a biased source distribution and severely degrade the performance of the domain adaptive detector de facto. In this paper, we represent the first effort to formulate noisy DAOD and propose a Noise Latent Transferability Exploration (NLTE) framework to address this issue. It is featured with 1) Potential Instance Mining (PIM), which leverages eligible proposals to recapture the miss-annotated instances from the background; 2) Morphable Graph Relation Module (MGRM), which models the adaptation feasibility and transition probability of nois

y samples with relation matrices; 3) Entropy-Aware Gradient Reconcilement (EAGR), which incorporates the semantic information into the discrimination process and enforces the gradients provided by noisy and clean samples to be consistent towards learning domain-invariant representations. A thorough evaluation on benchmark DAOD datasets with noisy source annotations validates the effectiveness of NLTE. In particular, NLTE improves the mAP by 8.4% under 60% corrupted annotations and even approaches the ideal upper bound of training on a clean source dataset.

****

Decoupled Multi-Task Learning With Cyclical Self-Regulation for Face Parsing
Qingping Zheng, Jiankang Deng, Zheng Zhu, Ying Li, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4156-4165

This paper probes intrinsic factors behind typical failure cases (e.g spatial inconsistency and boundary confusion) produced by the existing state-of-the-art method in face parsing. To tackle these problems, we propose a novel Decoupled Multi-task Learning with Cyclical Self-Regulation (DML-CSR) for face parsing. Specifically, DML-CSR designs a multi-task model which comprises face parsing, binary edge, and category edge detection. These tasks only share low-level encoder weights without high-level interactions between each other, enabling to decouple auxiliary modules from the whole network at the inference stage. To address spatial inconsistency, we develop a dynamic dual graph convolutional network to capture global contextual information without using any extra pooling operation. To handle boundary confusion in both single and multiple face scenarios, we exploit binary and category edge detection to jointly obtain generic geometric structure and fine-grained semantic clues of human faces. Besides, to prevent noisy labels from degrading model generalization during training, cyclical self-regulation is proposed to self-ensemble several model instances to get a new model and the resulting model then is used to self-distill subsequent models, through alternating iterations. Experiments show that our method achieves the new state-of-the-art performance on the Helen, CelebAMask-HQ, and Lapa datasets. The source code is available at https://github.com/deepinsight/insightface/tree/master/parsing/dml_csr.

****

Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer
Shuai Yang, Liming Jiang, Ziwei Liu, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7693-7702

Recent studies on StyleGAN show high performance on artistic portrait generation by transfer learning with limited data. In this paper, we explore more challenging exemplar-based high-resolution portrait style transfer by introducing a novel DualStyleGAN with flexible control of dual styles of the original face domain and the extended artistic portrait domain. Different from StyleGAN, DualStyleGAN provides a natural way of style transfer by characterizing the content and style of a portrait with an intrinsic style path and a new extrinsic style path, respectively. The delicately designed extrinsic style path enables our model to modulate both the color and complex structural styles hierarchically to precisely pastiche the style example. Furthermore, a novel progressive fine-tuning scheme is introduced to smoothly transform the generative space of the model to the target domain, even with the above modifications on the network architecture. Experiments demonstrate the superiority of DualStyleGAN over state-of-the-art methods in high-quality portrait style transfer and flexible style control.

****

Learning To Memorize Feature Hallucination for One-Shot Image Generation
Yu Xie, Yanwei Fu, Ying Tai, Yun Cao, Junwei Zhu, Chengjie Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9130-9139

This paper studies the task of One-Shot image Generation (OSG), where generation network learned on base dataset should be generalizable to synthesize images of novel categories with only one available sample per novel category. Most existi

ng methods for feature transfer in one-shot image generation only learn reusable features implicitly on pre-training tasks. Such methods would be likely to over fit pre-training tasks. In this paper, we propose a novel model to explicitly le arn and memorize reusable features that can help hallucinate novel category imag es. To be specific, our algorithm learns to decompose image features into the Ca tegory-Related (CR) and Category-Independent (CI) features. Our model learning t o memorize class-independent CI features which are further utilized by our featu re hallucination component to generate target novel category images. We validate our model on several benchmarks. Extensive experiments demonstrate that our mod el effectively boosts the OSG performance and can generate compelling and divers e samples.

****

AUV-Net: Learning Aligned UV Maps for Texture Transfer and Synthesis

Zhiqin Chen, Kangxue Yin, Sanja Fidler; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1465-1474

In this paper, we address the problem of texture representation for 3D shapes fo r the challenging and underexplored tasks of texture transfer and synthesis. Pre vious works either apply spherical texture maps which may lead to large distorti ons, or use continuous texture fields that yield smooth outputs lacking details. We argue that the traditional way of representing textures with images and link ing them to a 3D mesh via UV mapping is more desirable, since synthesizing 2D im ages is a well-studied problem. We propose AUV-Net which learns to embed 3D surf aces into a 2D aligned UV space, by mapping the corresponding semantic parts of different 3D shapes to the same location in the UV space. As a result, textures are aligned across objects, and can thus be easily synthesized by generative mod els of images. Texture alignment is learned in an unsupervised manner by a simpl e yet effective texture alignment module, taking inspiration from traditional wo rks on linear subspace learning. The learned UV mapping and aligned texture repr esentations enable a variety of applications including texture transfer, texture synthesis, and textured single view 3D reconstruction. We conduct experiments o n multiple datasets to demonstrate the effectiveness of our method.

****

Open-Vocabulary One-Stage Detection With Hierarchical Visual-Language Knowledge Distillation

Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhan g, Weiming Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 14074-14083

Open-vocabulary object detection aims to detect novel object categories beyond t he training set. The advanced open-vocabulary two-stage detectors employ instanc e-level visual-to-visual knowledge distillation to align the visual space of the detector with the semantic space of the Pre-trained Visual-Language Model (PVLM). However, in the more efficient one-stage detector, the absence of class-agnos tic object proposals hinders the knowledge distillation on unseen objects, leadi ng to severe performance degradation. In this paper, we propose a hierarchical v isual-language knowledge distillation method, i.e., HierKD, for open-vocabulary one-stage detection. Specifically, a global-level knowledge distillation is expl ored to transfer the knowledge of unseen categories from the PVLM to the detecto r. Moreover, we combine the proposed global-level knowledge distillation and the common instance-level knowledge distillation in a hierarchical structure to lea rn the knowledge of seen and unseen categories simultaneously. Extensive experim ents on MS-COCO show that our method significantly surpasses the previous best o ne-stage detector with 11.9% and 6.7% AP50 gains under the zero-shot detection a nd generalized zero-shot detection settings, and reduces the AP50 performance ga p from 14% to 7.3% compared to the best two-stage detector. Code will be publicl y available.

****

Glass: Geometric Latent Augmentation for Shape Spaces

Sanjeev Muralikrishnan, Siddhartha Chaudhuri, Noam Aigerman, Vladimir G. Kim, Ma tthew Fisher, Niloy J. Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18552-18561

We investigate the problem of training generative models on very sparse collecti ons of 3D models. Particularly, instead of using difficult-to-obtain large sets of 3D models, we demonstrate that geometrically-motivated energy functions can b e used to effectively augment and boost only a sparse collection of example (tra ining) models. Technically, we analyze the Hessian of the as-rigid-as-possible (ARAP) energy to adaptively sample from and project to the underlying (local) sha pe space, and use the augmented dataset to train a variational autoencoder (VAE) . We iterate the process, of building latent spaces of VAE and augmenting the as sociated dataset, to progressively reveal a richer and more expressive generativ e space for creating geometrically and semantically valid samples. We evaluate o ur method against a set of strong baselines, provide ablation studies, and demon strate application towards establishing shape correspondences. GLASS produces mu ltiple interesting and meaningful shape variations even when starting from as fe w as 3-10 training shapes. Our code is available at https: //sanjeevmk.github.io /glass_webpage/.

****

COAP: Compositional Articulated Occupancy of People

Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhöfer, Siyu Tang; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13201-13210

We present a novel neural implicit representation for articulated human bodies. Compared to explicit template meshes, neural implicit body representations provi de an efficient mechanism for modeling interactions with the environment, which is essential for human motion reconstruction and synthesis in 3D scenes. However , existing neural implicit bodies suffer from either poor generalization on high ly articulated poses or slow inference time. In this work, we observe that prior knowledge about the human body's shape and kinematic structure can be leveraged to improve generalization and efficiency. We decompose the full-body geometry i nto local body parts and employ a part-aware encoder-decoder architecture to lea rn neural articulated occupancy that models complex deformations locally. Our lo cal shape encoder represents the body deformation of not only the corresponding body part but also the neighboring body parts. The decoder incorporates the geom etric constraints of local body shape which significantly improves pose generali zation. We demonstrate that our model is suitable for resolving self-intersectio ns and collisions with 3D environments. Quantitative and qualitative experiments show that our method largely outperforms existing solutions in terms of both ef ficiency and accuracy.

****

Counterfactual Cycle-Consistent Learning for Instruction Following and Generatio n in Vision-Language Navigation

Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, Wenguan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 15471-15481

Since the rise of vision-language navigation (VLN), great progress has been made in instruction following -- building a follower to navigate environments under the guidance of instructions. However, far less attention has been paid to the i nverse task: instruction generation -- learning a speaker to generate grounded d escriptions for navigation routes. Existing VLN methods train a speaker independ ently and typically treat it as a data augmentation tool for strengthening the f ollower, while ignoring rich cross-task relations. Here we describe an approach that learns the two tasks simultaneously and exploits their intrinsic correlatio ns to boost the training of each: the follower judges whether the speaker-create d instruction explains the original navigation route correctly, and vice versa. Without the need of aligned instruction-path pairs, such cycle-consistent learni ng scheme is complementary to task-specific training objectives defined on label ed data, and can also be applied over unlabeled paths (sampled without paired in structions). Another agent, called creator, is added to generate counterfactual environments. It greatly changes current scenes yet leaves novel items -- which are crucial for the execution of original instructions -- unchanged. Thus more i nformative training scenes are synthesized and the three agents compose a powerf

ul VLN learning system. Experiments on a standard benchmark show that our approa ch improves the performance of various follower models and produces accurate nav igation instructions. Our code will be released.
****

Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solution s With Superior OOD Generalization
Damien Teney, Ehsan Abbasnejad, Simon Lucey, Anton van den Hengel; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 16761-16772
Neural networks trained with SGD were recently shown to rely preferentially on l inearly-predictive features and can ignore complex, equally-predictive ones. Thi s simplicity bias can explain their lack of robustness out of distribution (OOD) . The more complex the task to learn, the more likely it is that statistical art ifacts (i.e. selection biases, spurious correlations) are simpler than the mecha nisms to learn. We demonstrate that the simplicity bias can be mitigated and OOD generalization improved. We train a set of similar models to fit the data in di fferent ways using a penalty on the alignment of their input gradients. We show theoretically and empirically that this induces the learning of more complex pre dictive patterns. OOD generalization fundamentally requires information beyond i .i.d. examples, such as multiple training environments, counterfactual examples, or other side information. Our approach shows that we can defer this requiremen t to an independent model selection stage. We obtain SOTA results in visual reco gnition on biased data and generalization across visual domains. The method - th e first to evade the simplicity bias - highlights the need for a better understa nding and control of inductive biases in deep learning.
****

Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities
Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Ro bert Wang, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21096-21106
Assembly101 is a new procedural activity dataset featuring 4321 videos of people assembling and disassembling 101 "take-apart" toy vehicles. Participants work w ithout fixed instructions, and the sequences feature rich and natural variations in action ordering, mistakes, and corrections. Assembly101 is the first multi-v iew action dataset, with simultaneous static (8) and egocentric (4) recordings. Sequences are annotated with more than 100K coarse and 1M fine-grained action se gments, and 18M 3D hand poses. We benchmark on three action understanding tasks: recognition, anticipation and temporal segmentation. Additionally, we propose a novel task of detecting mistakes. The unique recording format and rich set of a nnotations allow us to investigate generalization to new toys, cross-view transf er, long-tailed distributions, and pose vs. appearance. We envision that Assembl y101 will serve as a new challenge to investigate various activity understanding problems.
****

Deterministic Point Cloud Registration via Novel Transformation Decomposition
Wen Chen, Haoang Li, Qiang Nie, Yun-Hui Liu; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6348-6356
Given a set of putative 3D-3D point correspondences, we aim to remove outliers a nd estimate rigid transformation with 6 degrees of freedom (DOF). Simultaneously estimating these 6 DOF is time-consuming due to high-dimensional parameter spac e. To solve this problem, it is common to decompose 6 DOF, i.e. independently co mpute 3-DOF rotation and 3-DOF translation. However, high non-linearity of 3-DOF rotation still limits the algorithm efficiency, especially when the number of c orrespondences is large. In contrast, we propose to decompose 6 DOF into (2+1) a nd (1+2) DOF. Specifically, (2+1) DOF represent 2-DOF rotation axis and 1-DOF di splacement along this rotation axis. (1+2) DOF indicate 1-DOF rotation angle and 2-DOF displacement orthogonal to the above rotation axis. To compute these DOF, we design a novel two-stage strategy based on inlier set maximization. By lever aging branch and bound, we first search for (2+1) DOF, and then the remaining (1

+2) DOF. Thanks to the proposed transformation decomposition and two-stage searc
h strategy, our method is deterministic and leads to low computational complexit
y. We extensively compare our method with state-of-the-art approaches. Our metho
d is more accurate and robust than the approaches that provide similar efficienc
y to ours. Our method is more efficient than the approaches whose accuracy and r
obustness are comparable to ours.
****

Motion-Adjustable Neural Implicit Video Representation
Long Mai, Feng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2022, pp. 10738-10747
Implicit neural representation (INR) has been successful in representing static
images. Contemporary image-based INR, with the use of Fourier-based positional e
ncoding, can be viewed as a mapping from sinusoidal patterns with different freq
uencies to image content. Inspired by that view, we hypothesize that it is possi
ble to generate temporally varying content with a single image-based INR model b
y displacing its input sinusoidal patterns over time. By exploiting the relation
 between the phase information in sinusoidal functions and their displacements,
we incorporate into the conventional image-based INR model a phase-varying posit
ional encoding module, and couple it with a phase-shift generation module that d
etermines the phase-shift values at each frame. The model is trained end-to-end
on a video to jointly determine the phase-shift values at each time with the map
ping from the phase-shifted sinusoidal functions to the corresponding frame, ena
bling an implicit video representation. Experiments on a wide range of videos su
ggest that such a model is capable of learning to interpret phase-varying positi
onal embeddings into the corresponding time-varying content. More importantly, w
e found that the learned phase-shift vectors tend to capture meaningful temporal
 and motion information from the video. In particular, manipulating the phase-sh
ift vectors induces meaningful changes in the temporal dynamics of the resulting
 video, enabling non-trivial temporal and motion editing effects such as tempora
l interpolation, motion magnification, motion smoothing, and video loop detectio
n.
****

Neural Prior for Trajectory Estimation
Chaoyang Wang, Xueqian Li, Jhony Kaesemodel Pontes, Simon Lucey; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
 pp. 6532-6542
Neural priors are a promising direction to capture low-level vision statistics w
ithout relying on handcrafted regularizers. Recent works have successfully shown
 the use of neural architecture biases to implicitly regularize image denoising,
 super-resolution, inpainting, synthesis, scene flow, among others. They do not
rely on large-scale datasets to capture prior statistics and thus generalize wel
l to out-of-the-distribution data. Inspired by such advances, we investigate neu
ral priors for trajectory representation. Traditionally, trajectories have been
represented by a set of handcrafted bases that have limited expressibility. Here
, we propose a neural trajectory prior to capture continuous spatio-temporal inf
ormation without the need for offline data. We demonstrate how our proposed obje
ctive is optimized during runtime to estimate trajectories for two important tas
ks: Non-Rigid Structure from Motion (NRSfM) and lidar scene flow integration for
 self-driving scenes. Our results are competitive to many state-of-the-art metho
ds for both tasks.
****

DPICT: Deep Progressive Image Compression Using Trit-Planes
Jae-Han Lee, Seungmin Jeon, Kwang Pyo Choi, Youngo Park, Chang-Su Kim; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 16113-16122
We propose the deep progressive image compression using trit-planes (DPICT) algo
rithm, which is the first learning-based codec supporting fine granular scalabil
ity (FGS). First, we transform an image into a latent tensor using an analysis n
etwork. Then, we represent the latent tensor in ternary digits (trits) and encod
e it into a compressed bitstream trit-plane by trit-plane in the decreasing orde

r of significance. Moreover, within each trit-plane, we sort the trits according to their rate-distortion priorities and transmit more important information first. Since the compression network is less optimized for the cases of using fewer trit-planes, we develop a postprocessing network for refining reconstructed images at low rates. Experimental results show that DPICT outperforms conventional progressive codecs significantly, while enabling FGS transmission. Codes are available at https://github.com/jaehanlee-mcl/DPICT.
****

Rethinking Depth Estimation for Multi-View Stereo: A Unified Representation
Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, Ronggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8645-8654
Depth estimation is solved as a regression or classification problem in existing learning-based multi-view stereo methods. Although these two representations have recently demonstrated their excellent performance, they still have apparent shortcomings, e.g., regression methods tend to overfit due to the indirect learning cost volume, and classification methods cannot directly infer the exact depth due to its discrete prediction. In this paper, we propose a novel representation, termed Unification, to unify the advantages of regression and classification. It can directly constrain the cost volume like classification methods, but also realize the sub-pixel depth prediction like regression methods. To excavate the potential of unification, we design a new loss function named Unified Focal Loss, which is more uniform and reasonable to combat the challenge of sample imbalance. Combining these two unburdened modules, we present a coarse-to-fine framework, that we call UniMVSNet. The results of ranking first on both DTU and Tanks and Temples benchmarks verify that our model not only performs the best but also has the best generalization ability.
****

Long-Tailed Recognition via Weight Balancing
Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, Shu Kong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6897-6907
In the real open world, data tends to follow long-tailed class distributions, motivating the well-studied long-tailed recognition (LTR) problem. Naive training produces models that are biased toward common classes in terms of higher accuracy. The key to addressing LTR is to balance various aspects including data distribution, training losses, and gradients in learning. We explore an orthogonal direction, weight balancing , motivated by the empirical observation that the naively trained classifier has "artificially" larger weights in norm for common classes (because there exists abundant data to train them, unlike the rare classes). We investigate three techniques to balance weights, L2-normalization, weight decay, and MaxNorm. We first point out that L2-normalization "perfectly" balances per-class weights to be unit norm, but such a hard constraint might prevent classes from learning better classifiers. In contrast, weight decay penalizes larger weights more heavily and so learns small balanced weights; the MaxNorm constraint encourages growing small weights within a norm ball but caps all the weights by the radius. Our extensive study shows that both help learn balanced weights and greatly improve the LTR accuracy. Surprisingly, weight decay, although underexplored in LTR, significantly improves over prior work. Therefore, we adopt a two-stage training paradigm and propose a simple approach to LTR: (1) learning features using the cross-entropy loss by tuning weight decay, and (2) learning classifiers using class-balanced loss by tuning weight decay and MaxNorm. Our approach achieves the state-of-the-art accuracy on five standard benchmarks, serving as a future baseline for long-tailed recognition.
****

Text to Image Generation With Semantic-Spatial Aware GAN
Wentong Liao, Kai Hu, Michael Ying Yang, Bodo Rosenhahn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18187-18196
Text-to-image synthesis (T2I) aims to generate photo-realistic images which are

semantically consistent with the text descriptions. Existing methods are usually built upon conditional generative adversarial networks (GANs) and initialize an image from noise with sentence embedding, and then refine the features with fine-grained word embedding iteratively. A close inspection of their generated images reveals a major limitation: even though the generated image holistically matches the description, individual image regions or parts of somethings are often not recognizable or consistent with words in the sentence, e.g. "a white crown". To address this problem, we propose a novel framework Semantic-Spatial Aware GAN for synthesizing images from input text. Concretely, we introduce a simple and effective Semantic-Spatial Aware block, which (1) learns semantic-adaptive transformation conditioned on text to effectively fuse text features and image features, and (2) learns a semantic mask in a weakly-supervised way that depends on the current text-image fusion process in order to guide the transformation spatially. Experiments on the challenging COCO and CUB bird datasets demonstrate the advantage of our method over the recent state-of-the-art approaches, regarding both visual fidelity and alignment with input text description. Code available at https://github.com/wtliao/text2image.

****

The Norm Must Go On: Dynamic Unsupervised Domain Adaptation by Normalization
M. Jehanzeb Mirza, Jakub Micorek, Horst Possegger, Horst Bischof; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14765-14775

Domain adaptation is crucial to adapt a learned model to new scenarios, such as domain shifts or changing data distributions. Current approaches usually require a large amount of labeled or unlabeled data from the shifted domain. This can be a hurdle in fields which require continuous dynamic adaptation or suffer from scarcity of data, e.g. autonomous driving in challenging weather conditions. To address this problem of continuous adaptation to distribution shifts, we propose Dynamic Unsupervised Adaptation (DUA). By continuously adapting the statistics of the batch normalization layers we modify the feature representations of the model. We show that by sequentially adapting a model with only a fraction of unlabeled data, a strong performance gain can be achieved. With even less than 1% of unlabeled data from the target domain, DUA already achieves competitive results to strong baselines. In addition, the computational overhead is minimal in contrast to previous approaches. Our approach is simple, yet effective and can be applied to any architecture which uses batch normalization as one of its components. We show the utility of DUA by evaluating it on a variety of domain adaptation datasets and tasks including object recognition, digit recognition and object detection.

****

ShapeFormer: Transformer-Based Shape Completion via Sparse Representation
Xingguang Yan, Liqiang Lin, Niloy J. Mitra, Dani Lischinski, Daniel Cohen-Or, Hui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6239-6249

We present ShapeFormer, a transformer-based network that produces a distribution of object completions, conditioned on incomplete, and possibly noisy, point clouds. The resultant distribution can then be sampled to generate likely completions, each of which exhibits plausible shape details, while being faithful to the input. To facilitate the use of transformers for 3D, we introduce a compact 3D representation, vector quantized deep implicit function (VQDIF), that utilizes spatial sparsity to represent a close approximation of a 3D shape by a short sequence of discrete variables. Experiments demonstrate that ShapeFormer outperforms prior art for shape completion from ambiguous partial inputs in terms of both completion quality and diversity. We also show that our approach effectively handles a variety of shape types, incomplete patterns, and real-world scans.

****

PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures
Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, Jacob Steinhardt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16783-16792

In real-world applications of machine learning, reliable and safe systems must c
onsider measures of performance beyond standard test set accuracy. These other g
oals include out-of-distribution (OOD) robustness, prediction consistency, resil
ience to adversaries, calibrated uncertainty estimates, and the ability to detec
t anomalous inputs. However, improving performance towards these goals is often
a balancing act that today's methods cannot achieve without sacrificing performa
nce on other safety axes. For instance, adversarial training improves adversaria
l robustness but sharply degrades other classifier performance metrics. Similarl
y, strong data augmentation and regularization techniques often improve OOD robu
stness but harm anomaly detection, raising the question of whether a Pareto impr
ovement on all existing safety measures is possible. To meet this challenge, we
design a new data augmentation strategy utilizing the natural structural complex
ity of pictures such as fractals, which outperforms numerous baselines, is near
Pareto-optimal, and roundly improves safety measures.
****

Eigencontours: Novel Contour Descriptors Based on Low-Rank Approximation
Wonhui Park, Dongkwon Jin, Chang-Su Kim; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2667-2675
Novel contour descriptors, called eigencontours, based on low-rank approximation
 are proposed in this paper. First, we construct a contour matrix containing all
 object boundaries in a training set. Second, we decompose the contour matrix in
to eigencontours via the best rank-M approximation. Third, we represent an objec
t boundary by a linear combination of the M eigencontours. We also incorporate t
he eigencontours into an instance segmentation framework. Experimental results d
emonstrate that the proposed eigencontours can represent object boundaries more
effectively and more efficiently than existing descriptors in a low-dimensional
space. Furthermore, the proposed algorithm yields meaningful performances on ins
tance segmentation datasets.
****

Generalizable Cross-Modality Medical Image Segmentation via Style Augmentation a
nd Dual Normalization
Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, Yinghuan Shi; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20856-20
865
For medical image segmentation, imagine if a model was only trained using MR ima
ges in source domain, how about its performance to directly segment CT images in
 target domain? This setting, namely generalizable cross-modality segmentation,
owning its clinical potential, is much more challenging than other related setti
ngs, e.g., domain adaptation. To achieve this goal, we in this paper propose a n
ovel dual-normalization model by leveraging the augmented source-similar and sou
rce-dissimilar images during our generalizable segmentation. To be specific, giv
en a single source domain, aiming to simulate the possible appearance change in
unseen target domains, we first utilize a nonlinear transformation to augment so
urce-similar and source-dissimilar images. Then, to sufficiently exploit these t
wo types of augmentations, our proposed dual-normalization based model employs a
 shared backbone yet independent batch normalization layer for separate normaliz
ation. Afterward, we put forward a style-based selection scheme to automatically
 choose the appropriate path in the test stage. Extensive experiments on three p
ublicly available datasets, i.e., BraTS, Cross-Modality Cardiac, and Abdominal M
ulti-Organ datasets, have demonstrated that our method outperforms other state-o
f-the-art domain generalization methods. Code is available at https://github.com
/zzzqzhou/Dual-Normalization.
****

Learning Optical Flow With Kernel Patch Attention
Ao Luo, Fan Yang, Xin Li, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8906-8915
Optical flow is a fundamental method used for quantitative motion estimation on
the image plane. In the deep learning era, most works treat it as a task of 'mat
ching of features', learning to pull matched pixels as close as possible in feat
ure space and vice versa. However, spatial affinity (smoothness constraint), ano

ther important component for motion understanding, has been largely overlooked. In this paper, we introduce a novel approach, called kernel patch attention (KPA), to better resolve the ambiguity in dense matching by explicitly taking the local context relations into consideration. Our KPA operates on each local patch, and learns to mine the context affinities for better inferring the flow fields. It can be plugged into contemporary optical flow architecture and empower the model to conduct comprehensive motion analysis with both feature similarities and spatial relations. On Sintel dataset, the proposed KPA-Flow achieves the best performance with EPE of 1.35 on clean pass and 2.36 on final pass, and it sets a new record of 4.60% in F1-all on KITTI-15 benchmark.
****

Learning To Prompt for Open-Vocabulary Object Detection With Vision-Language Model

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, Guoqi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14084-14093

Recently, vision-language pre-training shows great potential in open-vocabulary object detection, where detectors trained on base classes are devised for detecting new classes. The class text embedding is firstly generated by feeding prompts to the text encoder of a pre-trained vision-language model. It is then used as the region classifier to supervise the training of a detector. The key element that leads to the success of this model is the proper prompt, which requires careful words tuning and ingenious design. To avoid laborious prompt engineering, there are some prompt representation learning methods being proposed for the image classification task, which however can only be sub-optimal solutions when applied to the detection task. In this paper, we introduce a novel method, detection prompt (DetPro), to learn continuous prompt representations for open-vocabulary object detection based on the pre-trained vision-language model. Different from the previous classification-oriented methods, DetPro has two highlights: 1) a background interpretation scheme to include the proposals in image background into the prompt training; 2) a context grading scheme to separate proposals in image foreground for tailored prompt training. We assemble DetPro with ViLD, a recent state-of-the-art openworld object detector, and conduct experiments on the LVIS as well as transfer learning on the Pascal VOC, COCO, Objects365 datasets. Experimental results show that our DetPro outperforms the baseline ViLD [5] in all settings, e.g., +3.4 APbox and +3.0 APmask improvements on the novel classes of LVIS. Code and models are available at https://github.com/dyabel/detpro.
****

TimeReplayer: Unlocking the Potential of Event Cameras for Video Interpolation

Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, Jianxing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17804-17813

Recording fast motion in a high FPS (frame-per-second) requires expensive high-speed cameras. As an alternative, interpolating low-FPS videos from commodity cameras has attracted significant attention. If only low-FPS videos are available, motion assumptions (linear or quadratic) are necessary to infer intermediate frames, which fail to model complex motions. Event camera, a new camera with pixels producing events of brightness change at the temporal resolution of \mu s (10^-6 second), is a game-changing device to enable video interpolation at the presence of arbitrarily complex motion. Since event camera is a novel sensor, its potential has not been fulfilled due to the lack of processing algorithms. The pioneering work Time Lens introduced event cameras to video interpolation by designing optical devices to collect a large amount of paired training data of high-speed frames and events, which is too costly to scale. To fully unlock the potential of event cameras, this paper proposes a novel TimeReplayer algorithm to interpolate videos captured by commodity cameras with events. It is trained in an unsupervised cycle-consistent style, canceling the necessity of high-speed training data and bringing the additional ability of video extrapolation. Its state-of-the-art results and demo videos in supplementary reveal the promising future of event-based vision.

****
General Incremental Learning With Domain-Aware Categorical Representations
Jiangwei Xie, Shipeng Yan, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14351-14360
Continual learning is an important problem for achieving human-level intelligence in real-world applications as an agent must continuously accumulate knowledge in response to streaming data/tasks. In this work, we consider a general and yet under-explored incremental learning problem in which both the class distribution and class-specific domain distribution change over time. In addition to the typical challenges in class incremental learning, this setting also faces the intra-class stability-plasticity dilemma and intra-class domain imbalance problems. To address above issues, we develop a novel domain-aware continual learning method based on the EM framework. Specifically, we introduce a flexible class representation based on the von Mises-Fisher mixture model to capture the intra-class structure, using an expansion-and-reduction strategy to dynamically increase the number of components according to the class complexity. Moreover, we design a bi-level balanced memory to cope with data imbalances within and across classes, which combines with a distillation loss to achieve better inter- and intra-class stability-plasticity trade-off. We conduct exhaustive experiments on three benchmarks: iDigits, iDomainNet and iCIFAR-20. The results show that our approach consistently outperforms previous methods by a significant margin, demonstrating its superiority.
****
Interactive Segmentation and Visualization for Tiny Objects in Multi-Megapixel Images
Chengyuan Xu, Boning Dong, Noah Stier, Curtis McCully, D. Andrew Howell, Pradeep Sen, Tobias Höllerer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21447-21452
We introduce an interactive image segmentation and visualization framework for identifying, inspecting, and editing tiny objects (just a few pixels wide) in large multi-megapixel high-dynamic-range (HDR) images. Detecting cosmic rays (CRs) in astronomical observations is a cumbersome workflow that requires multiple tools, so we developed an interactive toolkit that unifies model inference, HDR image visualization, segmentation mask inspection and editing into a single graphical user interface. The feature set, initially designed for astronomical data, makes this work a useful research-supporting tool for human-in-the-loop tiny-object segmentation in scientific areas like biomedicine, materials science, remote sensing, etc., as well as computer vision. Our interface features mouse-controlled, synchronized, dual-window visualization of the image and the segmentation mask, a critical feature for locating tiny objects in multi-megapixel images. The browser-based tool can be readily hosted on the web to provide multi-user access and GPU acceleration for any device. The toolkit can also be used as a high-precision annotation tool, or adapted as the frontend for an interactive machine learning framework. Our open-source dataset, CR detection model, and visualization toolkit are available at https://github.com/cy-xu/cosmic-conn.
****
ActiveZero: Mixed Domain Learning for Active Stereovision With Zero Annotation
Isabella Liu, Edward Yang, Jianyu Tao, Rui Chen, Xiaoshuai Zhang, Qing Ran, Zhu Liu, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13033-13042
Traditional depth sensors generate accurate real world depth estimates that surpass even the most advanced learning approaches trained only on simulation domains. Since ground truth depth is readily available in the simulation domain but quite difficult to obtain in the real domain, we propose a method that leverages the best of both worlds. In this paper we present a new framework, ActiveZero, which is a mixed domain learning solution for active stereovision systems that requires no real world depth annotation. First, we demonstrate the transferability of our method to out-of-distribution real data by using a mixed domain learning strategy. In the simulation domain, we use a combination of supervised disparity loss and self-supervised losses on a shape primitives dataset. By contrast, in

the real domain, we only use self-supervised losses on a dataset that is out-of-distribution from either training simulation data or test real data. Second, our method introduces a novel self-supervised loss called temporal IR reprojection to increase the robustness and accuracy of our reprojections in hard-to-perceive regions. Finally, we show how the method can be trained end-to-end and that each module is important for attaining the end result. Extensive qualitative and quantitative evaluations on real data demonstrate state of the art results that can even beat a commercial depth sensor.
****

DearKD: Data-Efficient Early Knowledge Distillation for Vision Transformers
Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12052-12062
Transformers have been successfully applied to computer vision due to its powerful modelling capacity with self-attention. However, the good performance of transformers heavily depends on enormous training images. Thus, a data-efficient transformer solution is urgently needed. In this work, we propose an early knowledge distillation framework, which is termed as DearKD, to improvethe data-efficiency required by transformers. Our DearKD is a two-stage framework that first distills the inductive biases from the early intermediate layers of a CNN and then gives the transformer full play by training without distillation. Further, our DearKD can also be applied to the extreme data-free case where no real images are available, where we propose a boundary-preserving intra-divergence loss based on DeepInversion to further close the performance gap against the full-data counterpart. Extensive experiments on ImageNet, partial ImageNet, data-free setting and other downstream tasks prove the superiority of DearKD over its baselines and state-of-the-art methods.
****

Global-Aware Registration of Less-Overlap RGB-D Scans
Che Sun, Yunde Jia, Yi Guo, Yuwei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6357-6366
We propose a novel method of registering less-overlap RGB-D scans. Our method learns global information of a scene to construct a panorama, and aligns RGB-D scans to the panorama to perform registration. Different from existing methods that use local feature points to register less-overlap RGB-D scans and mismatch too much, we use global information to guide the registration, thereby alleviating the mismatching problem by preserving global consistency of alignments. To this end, we build a scene inference network to construct the panorama representing global information. We introduce a reinforcement learning strategy to iteratively align RGB-D scans with the panorama and refine the panorama representation, which reduces the noise of global information and preserves global consistency of both geometric and photometric alignments. Experimental results on benchmark datasets including SUNCG, Matterport, and ScanNet show the superiority of our method.
****

RayMVSNet: Learning Ray-Based 1D Implicit Fields for Accurate Multi-View Stereo
Junhua Xi, Yifei Shi, Yijie Wang, Yulan Guo, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8595-8605
Learning-based multi-view stereo (MVS) has by far centered around 3D convolution on cost volumes. Due to the high computation and memory consumption of 3D CNN, the resolution of output depth is often considerably limited. Different from most existing works dedicated to adaptive refinement of cost volumes, we opt to directly optimize the depth value along each camera ray, mimicking the range (depth) finding of a laser scanner. This reduces the MVS problem to ray-based depth optimization which is much more light-weight than full cost volume optimization. In particular, we propose RayMVSNet which learns sequential prediction of a 1D implicit field along each camera ray with the zero-crossing point indicating scene depth. This sequential modeling, conducted based on transformer features, essentially learns the epipolar line search in traditional multi-view stereo. We also devise a multi-task learning for better optimization convergence and depth accu

racy. Our method ranks top on both the DTU and the Tanks & Temples datasets over all previous learning-based methods, achieving overall reconstruction score of 0.33mm on DTU and f-score of 59.48% on Tanks & Temples.
****

ContrastMask: Contrastive Learning To Segment Every Thing
Xuehui Wang, Kai Zhao, Ruixin Zhang, Shouhong Ding, Yan Wang, Wei Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11604-11613
Partially-supervised instance segmentation is a task which requests segmenting objects from novel categories via learning on limited base categories with annotated masks thus eliminating demands of heavy annotation burden. The key to addressing this task is to build an effective class-agnostic mask segmentation model. Unlike previous methods that learn such models only on base categories, in this paper, we propose a new method, named ContrastMask, which learns a mask segmentation model on both base and novel categories under a unified pixel-level contrastive learning framework. In this framework, annotated masks of base categories and pseudo masks of novel categories serve as a prior for contrastive learning, where features from the mask regions (foreground) are pulled together, and are contrasted against those from the background, and vice versa. Through this framework, feature discrimination between foreground and background is largely improved, facilitating learning of the class-agnostic mask segmentation model. Exhaustive experiments on the COCO dataset demonstrate the superiority of our method, which outperforms previous state-of-the-arts.
****

Efficient Deep Embedded Subspace Clustering
Jinyu Cai, Jicong Fan, Wenzhong Guo, Shiping Wang, Yunhe Zhang, Zhao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1-10
Recently deep learning methods have shown significant progress in data clustering tasks. Deep clustering methods (including distance-based methods and subspace-based methods) integrate clustering and feature learning into a unified framework, where there is a mutual promotion between clustering and representation. However, deep subspace clustering methods are usually in the framework of self-expressive model and hence have quadratic time and space complexities, which prevents their applications in large-scale clustering and real-time clustering. In this paper, we propose a new mechanism for deep clustering. We aim to learn the subspace bases from deep representation in an iterative refining manner while the refined subspace bases help learning the representation of the deep neural networks in return. The proposed method is out of the self-expressive framework, scales to the sample size linearly, and is applicable to arbitrarily large datasets and online clustering scenarios. More importantly, the clustering accuracy of the proposed method is much higher than its competitors. Extensive comparison studies with state-of-the-art clustering approaches on benchmark datasets demonstrate the superiority of the proposed method.
****

Neural MoCon: Neural Motion Control for Physically Plausible Human Motion Capture
Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, Yangang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6417-6426
Due to the visual ambiguity, purely kinematic formulations on monocular human motion capture are often physically incorrect, biomechanically implausible, and can not reconstruct accurate interactions. In this work, we focus on exploiting the high-precision and non-differentiable physics simulator to incorporate dynamical constraints in motion capture. Our key-idea is to use real physical supervisions to train a target pose distribution prior for sampling-based motion control to capture physically plausible human motion. To obtain accurate reference motion with terrain interactions for the sampling, we first introduce an interaction constraint based on SDF (Signed Distance Field) to enforce appropriate ground contact modeling. We then design a novel two-branch decoder to avoid stochastic er

ror from pseudo ground-truth and train a distribution prior with the non-differe ntiable physics simulator. Finally, we regress the sampling distribution from th e current state of the physical character with the trained prior and sample sati sfied target poses to track the estimated reference motion. Qualitative and quan titative results show that we can obtain physically plausible human motion with complex terrain interactions, human shape variations, and diverse behaviors. Mor e information can be found at https://www.yangangwang.com/papers/HBZ-NM-2022-03. html
****

Revisiting Temporal Alignment for Video Restoration
Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, Jiangbo Lu; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 60 53-6062
Long-range temporal alignment is critical yet challenging for video restoration tasks. Recently, some works attempt to divide the long-range alignment into seve ral sub-alignments and handle them progressively. Although this operation is hel pful in modeling distant correspondences, error accumulation is inevitable due t o the propagation mechanism. In this work, we present a novel, generic iterative alignment module which employs a gradual refinement scheme for sub-alignments, yielding more accurate motion compensation. To further enhance the alignment acc uracy and temporal consistency, we develop a non-parametric re-weighting method, where the importance of each neighboring frame is adaptively evaluated in a spa tial-wise way for aggregation. By virtue of the proposed strategies, our model a chieves state-of-the-art performance on multiple benchmarks across a range of vi deo restoration tasks including video super-resolution, denoising and deblurring .
****

Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning
Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, Faisal Mahmood; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2022, pp. 16144-16155
Vision Transformers (ViTs) and their multi-scale and hierarchical variations hav e been successful at capturing image representations but their use has been gene rally studied for low-resolution images (e.g. - 256x256, 384x384). For gigapixel whole-slide imaging (WSI) in computational pathology, WSIs can be as large as 1 50000x150000 pixels at 20x magnification and exhibit a hierarchical structure of visual tokens across varying resolutions: from 16x16 images capture spatial pat terns among cells, to 4096x4096 images characterizing interactions within the ti ssue microenvironment. We introduce a new ViT architecture called the Hierarchic al Image Pyramid Transformer (HIPT), which leverages the natural hierarchical st ructure inherent in WSIs using two levels of self-supervised learning to learn h igh-resolution image representations. HIPT is pretrained across 33 cancer types using 10,678 gigapixel WSIs, 408,218 4096x4096 images, and 104M 256x256 images. We benchmark HIPT representations on 9 slide-level tasks, and demonstrate that: 1) HIPT with hierarchical pretraining outperforms current state-of-the-art metho ds for cancer subtyping and survival prediction, 2) self-supervised ViTs are abl e to model important inductive biases about the hierarchical structure of phenot ypes in the tumor microenvironment.
****

Neural Reflectance for Shape Recovery With Shadow Handling
Junxuan Li, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2022, pp. 16221-16230
This paper aims at recovering the shape of a scene with unknown, non-Lambertian, and possibly spatially-varying surface materials. When the shape of the object is highly complex and that shadows cast on the surface, the task becomes very ch allenging. To overcome these challenges, we propose a coordinate-based deep MLP (multilayer perceptron) to parameterize both the unknown 3D shape and the unknow n reflectance at every surface point. This network is able to leverage the obser ved photometric variance and shadows on the surface, and recover both surface sh

ape and general non-Lambertian reflectance. We explicitly predict cast shadows, mitigating possible artifacts on these shadowing regions, leading to higher estimation accuracy. Our framework is entirely self-supervised, in the sense that it requires neither ground truth shape nor BRDF. Tests on real-world images demonstrate that our method outperform existing methods by a significant margin. Thanks to the small size of the MLP-net, our method is an order of magnitude faster than previous CNN-based methods.
****

Rep-Net: Efficient On-Device Learning via Feature Reprogramming
Li Yang, Adnan Siraj Rakin, Deliang Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12277-12286
Transfer learning, where the goal is to transfer the well-trained deep learning models from a primary source task to a new task, is a crucial learning scheme for on-device machine learning, due to the fact that IoT/edge devices collect and then process massive data in our daily life. However, due to the tiny memory constraint in IoT/edge devices, such on-device learning requires ultra-small training memory footprint, bringing new challenges for memory-efficient learning. Many existing works solve this problem by reducing the number of trainable parameters. However, this doesn't directly translate to memory-saving since the major bottleneck is the activations, not parameters. To develop memory-efficient on-device transfer learning, in this work, we are the first to approach the concept of transfer learning from a new perspective of intermediate feature reprogramming of a pre-trained model (i.e., backbone). To perform this lightweight and memory-efficient reprogramming, we propose to train a tiny Reprogramming Network (Rep-Net) directly from the new task input data, while freezing the backbone model. The proposed Rep-Net model interchanges the features with the backbone model using an activation connector at regular intervals to mutually benefit both the backbone model and Rep-Net model features. Through extensive experiments, we validate each design specs of the proposed Rep-Net model in achieving highly memory-efficient on-device reprogramming. Our experiments establish the superior performance (i.e., low training memory and high accuracy) of Rep-Net compared to SOTA on-device transfer learning schemes across multiple benchmarks.
****

Surface Representation for Point Clouds
Haoxi Ran, Jun Liu, Chengjie Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18942-18952
Most prior work represents the shapes of point clouds by coordinates. However, it is insufficient to describe the local geometry directly. In this paper, we present RepSurf (representative surfaces), a novel representation of point clouds to explicitly depict the very local structure. We explore two variants of RepSurf, Triangular RepSurf and Umbrella RepSurf inspired by triangle meshes and umbrella curvature in computer graphics. We compute the representations of RepSurf by predefined geometric priors after surface reconstruction. RepSurf can be a plug-and-play module for most point cloud models thanks to its free collaboration with irregular points. Based on a simple baseline of PointNet++ (SSG version), Umbrella RepSurf surpasses the previous state-of-the-art by a large margin for classification, segmentation and detection on various benchmarks in terms of performance and efficiency. With an increase of around 0.008M number of parameters, 0.04G FLOPs, and 1.12ms inference time, our method achieves 94.7% (+0.5%) on ModelNet40, and 84.6% (+1.8%) on ScanObjectNN for classification, while 74.3% (+0.8%) mIoU on S3DIS 6-fold, and 70.0% (+1.6%) mIoU on ScanNet for segmentation. For detection, previous state-of-the-art detector with our RepSurf obtains 71.2% (+2.1%) mAP_25, 54.8% (+2.0%) mAP_50 on ScanNetV2, and 64.9% (+1.9%) mAP_25, 47.1% (+2.5%) mAP_50 on SUN RGB-D. Our lightweight Triangular RepSurf performs its excellence on these benchmarks as well. The code is publicly available at https://github.com/hancyran/RepSurf.
****

Implicit Motion Handling for Video Camouflaged Object Detection
Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, Zongyuan Ge; Proceedings of the IEEE/CVF Conference on Computer Vision an

d Pattern Recognition (CVPR), 2022, pp. 13864-13873
We propose a new video camouflaged object detection (VCOD) framework that can exploit both short-term dynamics and long-term temporal consistency to detect camouflaged objects from video frames. An essential property of camouflaged objects is that they usually exhibit patterns similar to the background and thus make them hard to identify from still images. Therefore, effectively handling temporal dynamics in videos becomes the key for the VCOD task as the camouflaged objects will be noticeable when they move. However, current VCOD methods often leverage homography or optical flows to represent motions, where the detection error may accumulate from both the motion estimation error and the segmentation error. On the other hand, our method unifies motion estimation and object segmentation within a single optimization framework. Specifically, we build a dense correlation volume to implicitly capture motions between neighbouring frames and utilize the final segmentation supervision to optimize the implicit motion estimation and segmentation jointly. Furthermore, to enforce temporal consistency within a video sequence, we jointly utilize a spatio-temporal transformer to refine the short-term predictions. Extensive experiments on VCOD benchmarks demonstrate the architectural effectiveness of our approach. We also provide a large-scale VCOD dataset named MoCA-Mask with pixel-level handcrafted ground-truth masks and construct a comprehensive VCOD benchmark with previous methods to facilitate research in this direction. Dataset Link: https://xueliancheng.github.io/SLT-Net-project.
****

OVE6D: Object Viewpoint Encoding for Depth-Based 6D Object Pose Estimation
Dingding Cai, Janne Heikkilä, Esa Rahtu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6803-6813
This paper proposes a universal framework, called OVE6D, for model-based 6D object pose estimation from a single depth image and a target object mask. Our model is trained using purely synthetic data rendered from ShapeNet, and, unlike most of the existing methods, it generalizes well on new real-world objects without any fine-tuning. We achieve this by decomposing the 6D pose into viewpoint, in-plane rotation around the camera optical axis and translation, and introducing novel lightweight modules for estimating each component in a cascaded manner. The resulting network contains less than 4M parameters while demonstrating excellent performance on the challenging T-LESS and Occluded LINEMOD datasets without any dataset-specific training. We show that OVE6D outperforms some contemporary deep learning-based pose estimation methods specifically trained for individual objects or datasets with real-world training data.
****

DeepLIIF: An Online Platform for Quantification of Clinical Pathology Slides
Parmida Ghahremani, Joseph Marino, Ricardo Dodds, Saad Nadeem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21399-21405
In the clinic, resected tissue samples are stained with Hematoxylin-and-Eosin (H&E) and/or Immunhistochemistry (IHC) stains and presented to the pathologists on glass slides or as digital scans for diagnosis and assessment of disease progression. Cell-level quantification, e.g. in IHC protein expression scoring, can be extremely inefficient and subjective. We present DeepLIIF (https://deepliif.org), a first free online platform for efficient and reproducible IHC scoring. DeepLIIF outperforms current state-of-the-art approaches (relying on manual error-prone annotations) by virtually restaining clinical IHC slides with more informative multiplex immunofluorescence staining. Our DeepLIIF cloud-native platform supports (1) more than 150 proprietary/non-proprietary input formats via the Bio-Formats standard, (2) interactive adjustment, visualization, and downloading of the IHC quantification results and the accompanying restained images, (3) consumption of an exposed workflow API programmatically or through interactive plugins for open source whole slide image viewers such as QuPath/ImageJ, and (4) auto scaling to efficiently scale GPU resources based on user demand.
****

Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer
Hao Jiang, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR), 2022, pp. 16388-16398
Video summarization has recently engaged increasing attention in computer vision communities. However, the scarcity of annotated data has been a key obstacle in this task. To address it, this work explores a new solution for video summarization by transferring samples from a correlated task (i.e., video moment localization) equipped with abundant training data. Our main insight is that the annotated video moments also indicate the semantic highlights of a video, essentially similar to video summary. Approximately, the video summary can be treated as a sparse, redundancy-free version of the video moments. Inspired by this observation, we propose an importance Propagation based collaborative Teaching Network (iPTNet). It consists of two separate modules that conduct video summarization and moment localization, respectively. Each module estimates a frame-wise importance map for indicating keyframes or moments. To perform cross-task sample transfer, we devise an importance propagation module that realizes the conversion between summarization-guided and localization-guided importance maps. This way critically enables optimizing one of the tasks using the data from the other task. Additionally, in order to avoid error amplification caused by batch-wise joint training, we devise a collaborative teaching scheme, which adopts a cross-task mean teaching strategy to realize the joint optimization of the two tasks and provide robust frame-level teaching signals. Extensive experiments on video summarization benchmarks demonstrate that iPTNet significantly outperforms previous state-of-the-art video summarization methods, serving as an effective solution that overcomes the data scarcity issue in video summarization.
****

WALT: Watch and Learn 2D Amodal Representation From Time-Lapse Imagery
N. Dinesh Reddy, Robert Tamburo, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9356-9366
Current methods for object detection, segmentation, and tracking fail in the presence of severe occlusions in busy urban environments. Labeled real data of occlusions is scarce (even in large datasets) and synthetic data leaves a domain gap, making it hard to explicitly model and learn occlusions. In this work, we present the best of both the real and synthetic worlds for automatic occlusion supervision using a large readily available source of data: time-lapse imagery from stationary webcams observing street intersections over weeks, months, or even years. We introduce a new dataset, Watch and Learn Time-lapse (WALT), consisting of 12 (4K and 1080p) cameras capturing urban environments over a year. We exploit this real data in a novel way to automatically mine a large set of unoccluded objects and then composite them in the same views to generate occlusions. This longitudinal self-supervision is strong enough for an amodal network to learn object-occluder-occluded layer representations. We show how to speed up the discovery of unoccluded objects and relate the confidence in this discovery to the rate and accuracy of training occluded objects. After watching and automatically learning for several days, this approach shows significant performance improvement in detecting and segmenting occluded people and vehicles, over human-supervised amodal approaches.
****

Learning With Twin Noisy Labels for Visible-Infrared Person Re-Identification
Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14308-14317
In this paper, we study an untouched problem in visible-infrared person re-identification (VI-ReID), namely, Twin Noise Labels (TNL) which refers to as noisy annotation and correspondence. In brief, on the one hand, it is inevitable to annotate some persons with the wrong identity due to the complexity in data collection and annotation, e.g., the poor recognizability in the infrared modality. On the other hand, the wrongly annotated data in a single modality will eventually contaminate the cross-modal correspondence, thus leading to noisy correspondence. To solve the TNL problem, we propose a novel method for robust VI-ReID, termed DuAlly Robust Training (DART). In brief, DART first computes the clean confidenc

e of annotations by resorting to the memorization effect of deep neural networks . Then, the proposed method rectifies the noisy correspondence with the estimate d confidence and further divides the data into four groups for further utilizati ons. Finally, DART employs a novel dually robust loss consisting of a soft ident ification loss and an adaptive quadruplet loss to achieve robustness on the nois y annotation and noisy correspondence. Extensive experiments on SYSU-MM01 and Re gDB datasets verify the effectiveness of our method against the twin noisy label s compared with five state-of-the-art methods. The code could be accessed from h ttps://github.com/XLearning-SCU/2022-CVPR-DART.
****

Optical Flow Estimation for Spiking Camera
Liwen Hu, Rui Zhao, Ziluo Ding, Lei Ma, Boxin Shi, Ruiqin Xiong, Tiejun Huang; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17844-17853
As a bio-inspired sensor with high temporal resolution, the spiking camera has a n enormous potential in real applications, especially for motion estimation in h igh-speed scenes. However, frame-based and event-based methods are not well suit ed to spike streams from the spiking camera due to the different data modalities . To this end, we present, SCFlow, a tailored deep learning pipeline to estimate optical flow in high-speed scenes from spike streams. Importantly, a novel inpu t representation is introduced which can adaptively remove the motion blur in sp ike streams according to the prior motion. Further, for training SCFlow, we synt hesize two sets of optical flow data for the spiking camera, SPIkingly Flying Th ings and Photo-realistic High-speed Motion, denoted as SPIFT and PHM respectivel y, corresponding to random high-speed and well-designed scenes. Experimental res ults show that the SCFlow can predict optical flow from spike streams in differe nt high-speed scenes. Moreover, SCFlow shows promising generalization on real sp ike streams. Codes and datasets refer to https://github.com/Acnext/Optical-Flow-For-Spiking-Camera.
****

MetaFormer Is Actually What You Need for Vision
Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng , Shuicheng Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2022, pp. 10819-10829
Transformers have shown great potential in computer vision tasks. A common belie f is their attention-based token mixer module contributes most to their competen ce. However, recent works show the attention-based module in transformers can be replaced by spatial MLPs and the resulted models still perform quite well. Base d on this observation, we hypothesize that the general architecture of the trans formers, instead of the specific token mixer module, is more essential to the mo del's performance. To verify this, we deliberately replace the attention module in transformers with an embarrassingly simple spatial pooling operator to conduc t only basic token mixing. Surprisingly, we observe that the derived model, term ed as PoolFormer, achieves competitive performance on multiple computer vision t asks. For example, on ImageNet-1K, PoolFormer achieves 82.1% top-1 accuracy, sur passing well-tuned vision transformer/MLP-like baselines DeiT-B/ResMLP-B24 by 0. 3%/1.1% accuracy with 35%/52% fewer parameters and 49%/61% fewer MACs. The effec tiveness of PoolFormer verifies our hypothesis and urges us to initiate the conc ept of "MetaFormer", a general architecture abstracted from transformers without specifying the token mixer. Based on the extensive experiments, we argue that M etaFormer is the key player in achieving superior results for recent transformer and MLP-like models on vision tasks. This work calls for more future research d edicated to improving MetaFormer instead of focusing on the token mixer modules. Additionally, our proposed PoolFormer could serve as a starting baseline for fu ture MetaFormer architecture design.
****

GradViT: Gradient Inversion of Vision Transformers
Ali Hatamizadeh, Hongxu Yin, Holger R. Roth, Wenqi Li, Jan Kautz, Daguang Xu, Pa vlo Molchanov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 10021-10030

In this work we demonstrate the vulnerability of vision transformers (ViTs) to g radient-based inversion attacks. During this attack, the original data batch is reconstructed given model weights and the corresponding gradients. We introduce a method, named GradViT, that optimizes random noise into naturally looking imag es via an iterative process. The optimization objective consists of (i) a loss o n matching the gradients, (ii) image prior in the form of distance to batch norm alization statistics of a pretrained CNN model, and (iii) a total variation regu larization on patches to guide correct recovery locations. We propose a unique l oss scheduling function to overcome local minima during optimization. We evaluat e GadViT on ImageNet1K and MS-Celeb-1M datasets, and observe unprecedentedly hig h fidelity and closeness to the original (hidden) data. During the analysis we f ind that vision transformers are significantly more vulnerable than previously s tudied CNNs due to the presence of the attention mechanism. Our method demonstra tes new state-of-the-art results for gradient inversion in both qualitative and quantitative metrics. Project page at https://gradvit.github.io.
****

Spatial-Temporal Space Hand-in-Hand: Spatial-Temporal Video Super-Resolution via Cycle-Projected Mutual Learning
Mengshun Hu, Kui Jiang, Liang Liao, Jing Xiao, Junjun Jiang, Zheng Wang; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3574-3583
Spatial-Temporal Video Super-Resolution (ST-VSR) aims to generate super-resolved videos with higher resolution (HR) and higher frame rate (HFR). Quite intuitive ly, pioneering two-stage based methods complete ST-VSR directly combining two su b-tasks: Spatial Video Super-Resolution (S-VSR) and Temporal Video Super-Resolut ion (T-VSR) but ignore the reciprocal relations among them. Specifically, 1) T-V SR to S-VSR: temporal correlations help accurate spatial detail representation w ith more clues; 2) S-VSR to T-VSR: abundant spatial information contributes to t he refinement of temporal prediction. To this end, we propose a one-stage based Cycle-projected Mutual learning network (CycMu-Net) for ST-VSR, which makes full use of spatial-temporal correlations via the mutual learning between S-VSR and T-VSR. Specifically, we propose to exploit the mutual information among them via iterative up-and-down projections, where the spatial and temporal features are fully fused and distilled, helping the high-quality video reconstruction. Beside s extensive experiments on benchmark datasets, we also compare our proposed CycM u-Net with S-VSR and T-VSR tasks, demonstrating that our method significantly ou tperforms state-of-the-art methods. Codes are publicly available at: https://git hub.com/hhhhhumengshun/CycMuNet.
****

InstaFormer: Instance-Aware Image-to-Image Translation With Transformer
Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, Seungryong Kim; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18321-18331
We present a novel Transformer-based network architecture for instance-aware ima ge-to-image translation, dubbed InstaFormer, to effectively integrate global- an d instance-level information. By considering extracted content features from an image as tokens, our networks discover global consensus of content features by c onsidering context information through a self-attention module in Transformers. By augmenting such tokens with an instance-level feature extracted from the cont ent feature with respect to bounding box information, our framework is capable o f learning an interaction between object instances and the global image, thus bo osting the instance-awareness. We replace layer normalization (LayerNorm) in sta ndard Transformers with adaptive instance normalization (AdaIN) to enable a mult i-modal translation with style codes. In addition, to improve the instance-aware ness and translation quality at object regions, we present an instance-level con tent contrastive loss defined between input and translated image. We conduct exp eriments to demonstrate the effectiveness of our InstaFormer over the latest met hods and provide extensive ablation studies.
****

Revisiting Near/Remote Sensing With Geospatial Attention

Scott Workman, M. Usman Rafique, Hunter Blanton, Nathan Jacobs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1778-1787

This work addresses the task of overhead image segmentation when auxiliary ground-level images are available. Recent work has shown that performing joint inference over these two modalities, often called near/remote sensing, can yield significant accuracy improvements. Extending this line of work, we introduce the concept of geospatial attention, a geometry-aware attention mechanism that explicitly considers the geospatial relationship between the pixels in a ground-level image and a geographic location. We propose an approach for computing geospatial attention that incorporates geometric features and the appearance of the overhead and ground-level imagery. We introduce a novel architecture for near/remote sensing that is based on geospatial attention and demonstrate its use for five segmentation tasks. The results demonstrate that our method significantly outperforms the previous state-of-the-art methods.

****

Joint Global and Local Hierarchical Priors for Learned Image Compression

Jun-Hyuk Kim, Byeongho Heo, Jong-Seok Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5992-6001

Recently, learned image compression methods have outperformed traditional hand-crafted ones including BPG. One of the keys to this success is learned entropy models that estimate the probability distribution of the quantized latent representation. Like other vision tasks, most recent learned entropy models are based on convolutional neural networks (CNNs). However, CNNs have a limitation in modeling long-range dependencies due to their nature of local connectivity, which can be a significant bottleneck in image compression where reducing spatial redundancy is a key point. To overcome this issue, we propose a novel entropy model called Information Transformer (Informer) that exploits both global and local information in a content-dependent manner using an attention mechanism. Our experiments show that Informer improves rate-distortion performance over the state-of-the-art methods on the Kodak and Tecnick datasets without the quadratic computational complexity problem. Our source code is available at https://github.com/naver-ai/informer.

****

Knowledge Distillation via the Target-Aware Transformer

Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, Gang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10915-10924

Knowledge distillation becomes a de facto standard to improve the performance of small neural networks. Most of the previous works propose to regress the representational features from the teacher to the student in a one-to-one spatial matching fashion. However, people tend to overlook the fact that, due to the architecture differences, the semantic information on the same spatial location usually vary. This greatly undermines the underlying assumption of the one-to-one distillation approach. To this end, we propose a novel one-to-all spatial matching knowledge distillation approach. Specifically, we allow each pixel of the teacher feature to be distilled to all spatial locations of the student features given its similarity, which is generated from a target-aware transformer. Our approach surpasses the state-of-the-art methods by a significant margin on various computer vision benchmarks, such as ImageNet, Pascal VOC and COCOStuff10k. Code is available at https://github.com/sihaoevery/TaT.

****

Recurring the Transformer for Video Action Recognition

Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, Dahai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14063-14073

Existing video understanding approaches, such as 3D convolutional neural networks and Transformer-Based methods, usually process the videos in a clip-wise manner. Hence huge GPU memory is needed, and fixed-length video clips are usually required. We introduce a novel Recurrent Vision Transformer (RViT) framework for sp

atial-temporal representation learning to achieve the video action recognition t ask. Specifically, the proposed RViT is equipped with an attention gate which is utilized to build interaction between current frame input and previous hidden s tate, thus aggregating the global level inter-frame features through the hidden state. RViT is executed recurrently to process a video clip by giving the curren t frame and previous hidden state. The RViT can capture both spatial and tempora l features because of the attention gate and recurrent execution. Besides, the p roposed RViT can work on both fixed-length and variant-length video clips proper ly without requiring large GPU memory thanks to the frame by frame processing fl ow. Our experiment results verify that RViT can achieve state-of-the-art perform ance on various datasets for the video recognition task. Specifically, RViT can achieve a top-1 accuracy of 81.5% on Kinetics-400, 92.31% on Jester, 67.9% on So mething-Something-V2, and an mAP accuracy of 66.1% on Charades.

****

Subspace Adversarial Training

Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, Xiaolin Huang; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 134 09-13418

Single-step adversarial training (AT) has received wide attention as it proved t o be both efficient and robust. However, a serious problem of catastrophic overf itting exists, i.e., the robust accuracy against projected gradient descent (PGD) attack suddenly drops to 0% during the training. In this paper, we approach th is problem from a novel perspective of optimization and firstly reveal the close link between the fast-growing gradient of each sample and overfitting, which ca n also be applied to understand robust overfitting in multi-step AT. To control the growth of the gradient, we propose a new AT method, Subspace Adversarial Tra ining (Sub-AT), which constrains AT in a carefully extracted subspace. It succes sfully resolves both kinds of overfitting and significantly boosts the robustnes s. In subspace, we also allow single-step AT with larger steps and larger radius , further improving the robustness performance. As a result, we achieve state-of -the-art single-step AT performance. Without any regularization term, our single -step AT can reach over 51% robust accuracy against strong PGD-50 attack of radi us 8/255 on CIFAR-10, reaching a competitive performance against standard multi- step PGD-10 AT with huge computational advantages. The code is released at https ://github.com/nblt/Sub-AT.

****

3D-VField: Adversarial Augmentation of Point Clouds for Domain Generalization in 3D Object Detection

Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Moha mmad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, Federico Tombari; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17295-17304

As 3D object detection on point clouds relies on the geometrical relationships b etween the points, non-standard object shapes can hinder a method's detection ca pability. However, in safety-critical settings, robustness to out-of-domain and long-tail samples is fundamental to circumvent dangerous issues, such as the mis detection of damaged or rare cars. In this work, we substantially improve the ge neralization of 3D object detectors to out-of-domain data by deforming point clo uds during training. We achieve this with 3D-VField: a novel data augmentation m ethod that plausibly deforms objects via vector fields learned in an adversarial fashion. Our approach constrains 3D points to slide along their sensor view ray s while neither adding nor removing any of them. The obtained vectors are transf erable, sample-independent and preserve shape and occlusions. Despite training o nly on a standard dataset, such as KITTI, augmenting with our vector fields sign ificantly improves the generalization to differently shaped objects and scenes. Towards this end, we propose and share CrashD: a synthetic dataset of realistic damaged and rare cars, with a variety of crash scenarios. Extensive experiments on KITTI, Waymo, our CrashD and SUN RGB-D show the generalizability of our techn iques to out-of-domain data, different models and sensors, namely LiDAR and ToF cameras, for both indoor and outdoor scenes. Our CrashD dataset is available at

https://crashd-cars.github.io.
****
Image Segmentation Using Text and Image Prompts
Timo Lüddecke, Alexander Ecker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7086-7096
Image segmentation is usually addressed by training a model for a fixed set of object classes. Incorporating additional classes or more complex queries later is expensive as it requires re-training the model on a dataset that encompasses these expressions. Here we propose a system that can generate image segmentations based on arbitrary prompts at test time. A prompt can be either a text or an image. This approach enables us to create a unified model (trained once) for three common segmentation tasks, which come with distinct challenges: referring expression segmentation, zero-shot segmentation and one-shot segmentation. We build up on the CLIP model as a backbone which we extend with a transformer-based decoder that enables dense prediction. After training on an extended version of the PhraseCut dataset, our system generates a binary segmentation map for an image based on a free-text prompt or on an additional image expressing the query. We analyze different variants of the latter image-based prompts in detail. This novel hybrid input allows for dynamic adaptation not only to the three segmentation tasks mentioned above, but to any binary segmentation task where a text or image query can be formulated. Finally, we find our system to adapt well to generalized queries involving affordances or properties. Code is available at https://eckerlab.org/code/clipseg
****
AutoMine: An Unmanned Mine Dataset
Yuchen Li, Zixuan Li, Siyu Teng, Yu Zhang, Yuhang Zhou, Yuchang Zhu, Dongpu Cao, Bin Tian, Yunfeng Ai, Zhe Xuanyuan, Long Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21308-21317
Autonomous driving datasets have played an important role in validating the advancement of intelligent vehicle algorithms including localization, perception and prediction in academic areas. However, current existing datasets pay more attention to the structured urban road, which hampers the exploration on unstructured special scenarios. Moreover, the open-pit mine is one of the typical representatives for them. Therefore, we introduce the Autonomous driving dataset on the Mining scene (AutoMine) for positioning and perception tasks in this paper. The AutoMine is collected by multiple acquisition platforms including an SUV, a wide-body mining truck and an ordinary mining truck, depending on the actual mine operation scenarios. The dataset consists of 18+ driving hours, 18K annotated lidar and image frames for 3D perception with various mines, time-of-the-day and weather conditions. The main contributions of the AutoMine dataset are as follows: 1. The first autonomous driving dataset for perception and localization in mine scenarios. 2.There are abundant dynamic obstacles of 9 degrees of freedom with large dimension difference (mining trucks and pedestrians) and extreme climatic conditions (the dust and snow) in the mining area. 3.Multi-platform acquisition strategies could capture mining data from multiple perspectives that fit the actual operation. More details can be found in our website(https://automine.cc).
****
Neural Data-Dependent Transform for Learned Image Compression
Dezhao Wang, Wenhan Yang, Yueyu Hu, Jiaying Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17379-17388
Learned image compression has achieved great success due to its excellent modeling capacity, but seldom further considers the Rate-Distortion Optimization (RDO) of each input image. To explore this potential in the learned codec, we make the first attempt to build a neural data-dependent transform and introduce a continuous online mode decision mechanism to jointly optimize the coding efficiency for each individual image. Specifically, apart from the image content stream, we employ an additional model stream to generate the transform parameters at the decoder side. The presence of a model stream enables our model to learn more abstract neural-syntax, which helps cluster the latent representations of images more compactly. Beyond the transform stage, we also adopt neural-syntax based post-p

rocessing for the scenarios that require higher quality reconstructions regardless of extra decoding overhead. Moreover, the involvement of the model stream further makes it possible to optimize both the representation and the decoder in an online way, i.e. RDO at the testing time. It is equivalent to a continuous online mode decision, like coding modes in the traditional codecs, to improve the coding efficiency based on the individual input image. The experimental results show the effectiveness of the proposed neural-syntax design and the continuous online mode decision mechanism, demonstrating the superiority of our method in coding efficiency.

****

Background Activation Suppression for Weakly Supervised Object Localization
Pingyu Wu, Wei Zhai, Yang Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14248-14257
Weakly supervised object localization (WSOL) aims to localize objects using only image-level labels. Recently a new paradigm has emerged by generating a foreground prediction map (FPM) to achieve localization task. Existing FPM-based methods use cross-entropy (CE) to evaluate the foreground prediction map and to guide the learning of generator. We argue for using activation value to achieve more efficient learning. It is based on the experimental observation that, for a trained network, CE converges to zero when the foreground mask covers only part of the object region. While activation value increases until the mask expands to the object boundary, which indicates that more object areas can be learned by using activation value. In this paper, we propose a Background Activation Suppression (BAS) method. Specifically, an Activation Map Constraint module (AMC) is designed to facilitate the learning of generator by suppressing the background activation value. Meanwhile, by using the foreground region guidance and the area constraint, BAS can learn the whole region of the object. In the inference phase, we consider the prediction maps of different categories together to obtain the final localization results. Extensive experiments show that BAS achieves significant and consistent improvement over the baseline methods on the CUB-200-2011 and ILSVRC datasets. Code and models are available at https://github.com/wpy1999/BAS.

****

How Many Observations Are Enough? Knowledge Distillation for Trajectory Forecasting
Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, Rita Cucchiara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6553-6562
Accurate prediction of future human positions is an essential task for modern video-surveillance systems. Current state-of-the-art models usually rely on a "history" of past tracked locations (e.g., 3 to 5 seconds) to predict a plausible sequence of future locations (e.g., up to the next 5 seconds). We feel that this common schema neglects critical traits of realistic applications: as the collection of input trajectories involves machine perception (i.e., detection and tracking), incorrect detection and fragmentation errors may accumulate in crowded scenes, leading to tracking drifts. On this account, the model would be fed with corrupted and noisy input data, thus fatally affecting its prediction performance. In this regard, we focus on delivering accurate predictions when only a few input observations are used, thus potentially lowering the risks associated with automatic perception. To this end, we conceive a novel distillation strategy that allows a knowledge transfer from a teacher network to a student one, the latter fed with fewer observations (just two ones). We show that a properly defined teacher supervision allows a student network to perform comparably to state-of-the-art approaches that demand more observations. Besides, extensive experiments on common trajectory forecasting datasets highlight that our student network better generalizes to unseen scenarios.

****

Evaluation-Oriented Knowledge Distillation for Deep Face Recognition
Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18740-18749

Knowledge distillation (KD) is a widely-used technique that utilizes large networks to improve the performance of compact models. Previous KD approaches usually aim to guide the student to mimic the teacher's behavior completely in the representation space. However, such one-to-one corresponding constraints may lead to inflexible knowledge transfer from the teacher to the student, especially those with low model capacities. Inspired by the ultimate goal of KD methods, we propose a novel Evaluation oriented KD method (EKD) for deep face recognition to directly reduce the performance gap between the teacher and student models during training. Specifically, we adopt the commonly used evaluation metrics in face recognition, i.e., False Positive Rate (FPR) and True Positive Rate (TPR) as the performance indicator. According to the evaluation protocol, the critical pair relations that cause the TPR and FPR difference between the teacher and student models are selected. Then, the critical relations in the student are constrained to approximate the corresponding ones in the teacher by a novel rank-based loss function, giving more flexibility to the student with low capacity. Extensive experimental results on popular benchmarks demonstrate the superiority of our EKD over state-of-the-art competitors.

****

Improving Subgraph Recognition With Variational Graph Information Bottleneck
Junchi Yu, Jie Cao, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19396-19405

Subgraph recognition aims at discovering a compressed substructure of a graph that is most informative to the graph property. It can be formulated by optimizing Graph Information Bottleneck (GIB) with a mutual information estimator. However, GIB suffers from training instability and degenerated results due to its intrinsic optimization process. To tackle these issues, we reformulate the subgraph recognition problem into two steps: graph perturbation and subgraph selection, leading to a novel Variational Graph Information Bottleneck (VGIB) framework. VGIB first employs the noise injection to modulate the information flow from the input graph to the perturbed graph. Then, the perturbed graph is encouraged to be informative to the graph property. VGIB further obtains the desired subgraph by filtering out the noise in the perturbed graph. With the customized noise prior for each input, the VGIB objective is endowed with a tractable variational upper bound, leading to a superior empirical performance as well as theoretical properties. Extensive experiments on graph interpretation, explainability of Graph Neural Networks, and graph classification show that VGIB finds better subgraphs than existing methods Extensive experiments on the explainability of Graph Neural Networks, graph interpretation, and graph classification show that VGIB finds better subgraphs than existing methods.

****

Slot-VPS: Object-Centric Representation Learning for Video Panoptic Segmentation
Yi Zhou, Hui Zhang, Hana Lee, Shuyang Sun, Pingjun Li, Yangguang Zhu, ByungIn Yoo, Xiaojuan Qi, Jae-Joon Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3093-3103

Video Panoptic Segmentation (VPS) aims at assigning a class label to each pixel, uniquely segmenting and identifying all object instances consistently across all frames. Classic solutions usually decompose the VPS task into several sub-tasks and utilize multiple surrogates (e.g. boxes and masks, centers and offsets) to represent objects. However, this divide-and-conquer strategy requires complex post-processing in both spatial and temporal domains and is vulnerable to failures from surrogate tasks. In this paper, inspired by object-centric learning which learns compact and robust object representations, we present Slot-VPS, the first end-to-end framework for this task. We encode all panoptic entities in a video, including both foreground instances and background semantics, in a unified representation called panoptic slots. The coherent spatio-temporal object's information is retrieved and encoded into the panoptic slots by the proposed Video Panoptic Retriever, enabling to localize, segment, differentiate, and associate objects in a unified manner. Finally, the output panoptic slots can be directly converted into the class, mask, and object ID of panoptic objects in the video. We conduct extensive ablation studies and demonstrate the effectiveness of our appro

ach on two benchmark datasets, Cityscapes-VPS (val and test sets) and VIPER (val set), achieving new state-of-the-art performance of 63.7, 63.3 and 56.2 VPQ, respectively.
****

Motion-From-Blur: 3D Shape and Motion Estimation of Motion-Blurred Objects in Videos
Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15990-15999
We propose a method for jointly estimating the 3D motion, 3D shape, and appearance of highly motion-blurred objects from a video. To this end, we model the blurred appearance of a fast moving object in a generative fashion by parametrizing its 3D position, rotation, velocity, acceleration, bounces, shape, and texture over the duration of a predefined time window spanning multiple frames. Using differentiable rendering, we are able to estimate all parameters by minimizing the pixel-wise reprojection error to the input video via backpropagating through a rendering pipeline that accounts for motion blur by averaging the graphics output over short time intervals. For that purpose, we also estimate the camera exposure gap time within the same optimization. To account for abrupt motion changes like bounces, we model the motion trajectory as a piece-wise polynomial, and we are able to estimate the specific time of the bounce at sub-frame accuracy. Experiments on established benchmark datasets demonstrate that our method outperforms previous methods for fast moving object deblurring and 3D reconstruction.
****

Efficient Video Instance Segmentation via Tracklet Query and Proposal
Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, Gérard Medioni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 959-968
Video Instance Segmentation (VIS) aims to simultaneously classify, segment, and track multiple object instances in videos. Recent clip-level VIS takes a short video clip as input each time showing stronger performance than frame-level VIS (tracking-by-segmentation), as more temporal context from multiple frames is utilized. Yet, most clip-level methods are neither end-to-end learnable nor real-time. These limitations are addressed by the recent VIS transformer (VisTR) which performs VIS end-to-end within a clip. However, VisTR suffers from long training time due to its frame-wise dense attention. In addition, VisTR is not fully end-to-end learnable in multiple video clips as it requires a hand-crafted data association to link instance tracklets between successive clips. This paper proposes EfficientVIS, a fully end-to-end framework with efficient training and inference. At the core are tracklet query and tracklet proposal that associate and segment regions-of-interest (RoIs) across space and time by an iterative query-video interaction. We further propose a correspondence learning that makes tracklets linking between clips end-to-end learnable. Compared to VisTR, EfficientVIS requires 15x fewer training epochs while achieving state-of-the-art accuracy on the YouTube-VIS benchmark. Meanwhile, our method enables whole video instance segmentation in a single end-to-end pass without data association at all.
****

Synthetic Generation of Face Videos With Plethysmograph Physiology
Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20587-20596
Accelerated by telemedicine, advances in Remote Photoplethysmography (rPPG) are beginning to offer a viable path toward non-contact physiological measurement. Unfortunately, the datasets for rPPG are limited as they require videos of the human face paired with ground-truth, synchronized heart rate data from a medical-grade health monitor. Also troubling is that the datasets are not inclusive of diverse populations, i.e., current real rPPG facial video datasets are imbalanced in terms of races or skin tones, leading to accuracy disparities on different demographic groups. This paper proposes a scalable biophysical learning based meth

od to generate physio-realistic synthetic rPPG videos given any reference image and target rPPG signal and shows that it could further improve the state-of-the-art physiological measurement and reduce the bias among different groups. We also collect the largest rPPG dataset of its kind (UCLA-rPPG) with a diverse presence of subject skin tones, in the hope that this could serve as a benchmark dataset for different skin tones in this area and ensure that advances of the technique can benefit all people for healthcare equity. The dataset is available at https://visual.ee.ucla.edu/rppg_avatars.htm/.
****

TransRAC: Encoding Multi-Scale Temporal Correlation With Transformers for Repetitive Action Counting
Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19013-19022
Counting repetitive actions are widely seen in human activities such as physical exercise. Existing methods focus on performing repetitive action counting in short videos, which is tough for dealing with longer videos in more realistic scenarios. In the data-driven era, the degradation of such generalization capability is mainly attributed to the lack of long video datasets. To complement this margin, we introduce a new large-scale repetitive action counting dataset covering a wide variety of video lengths, along with more realistic situations where action interruption or action inconsistencies occur in the video. Besides, we also provide a fine-grained annotation of the action cycles instead of just counting annotation along with a numerical value. Such a dataset contains 1,451 videos with about 20,000 annotations, which is more challenging. For repetitive action counting towards more realistic scenarios, we further propose encoding multi-scale temporal correlation with transformers that can take into account both performance and efficiency. Furthermore, with the help of fine-grained annotation of action cycles, we propose a density map regression-based method to predict the action period, which yields better performance with sufficient interpretability. Our proposed method outperforms state-of-the-art methods on all datasets and also achieves better performance on the unseen dataset without fine-tuning. The dataset and code are available.
****

Hallucinated Neural Radiance Fields in the Wild
Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, Jue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12943-12952
Neural Radiance Fields (NeRF) has recently gained popularity for its impressive novel view synthesis ability. This paper studies the problem of hallucinated NeRF: i.e., recovering a realistic NeRF at a different time of day from a group of tourism images. Existing solutions adopt NeRF with a controllable appearance embedding to render novel views under various conditions, but they cannot render view-consistent images with an unseen appearance. To solve this problem, we present an end-to-end framework for constructing a hallucinated NeRF, dubbed as Ha-NeRF. Specifically, we propose an appearance hallucination module to handle time-varying appearances and transfer them to novel views. Considering the complex occlusions of tourism images, we introduce an anti-occlusion module to decompose the static subjects for visibility accurately. Experimental results on synthetic data and real tourism photo collections demonstrate that our method can hallucinate the desired appearances and render occlusion-free images from different views. The project and supplementary materials are available at https://rover-xingyu.github.io/Ha-NeRF/.
****

NeuralHDHair: Automatic High-Fidelity Hair Modeling From a Single Image Using Implicit Neural Representations
Keyu Wu, Yifan Ye, Lingchen Yang, Hongbo Fu, Kun Zhou, Youyi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1526-1535
Undoubtedly, high-fidelity 3D hair plays an indispensable role in digital humans

. However, existing monocular hair modeling methods are either tricky to deploy in digital systems (e.g., due to their dependence on complex user interactions o r large databases) or can produce only a coarse geometry. In this paper, we intr oduce NeuralHDHair, a flexible, fully automatic system for modeling high-fidelit y hair from a single image. The key enablers of our system are two carefully des igned neural networks: an IRHairNet (Implicit representation for hair using neur al network) for inferring high-fidelity 3D hair geometric features (3D orientati on field and 3D occupancy field) hierarchically and a GrowingNet (Growing hair s trands using neural network) to efficiently generate 3D hair strands in parallel . Specifically, we perform a coarse-to-fine manner and propose a novel voxel-ali gned implicit function (VIFu) to represent the global hair feature, which is fur ther enhanced by the local details extracted from a hair luminance map. To impro ve the efficiency of a traditional hair growth algorithm, we adopt a local neura l implicit function to grow strands based on the estimated 3D hair geometric fea tures. Extensive experiments show that our method is capable of constructing a h igh-fidelity 3D hair model from a single image, both efficiently and effectively , and achieves the-state-of-the-art performance.

****

The Two Dimensions of Worst-Case Training and Their Integrated Effect for Out-of -Domain Generalization

Zeyi Huang, Haohan Wang, Dong Huang, Yong Jae Lee, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9631-9641

Training with an emphasis on "hard-to-learn" components of the data has been pro ven as an effective method to improve the generalization of machine learning mod els, especially in the settings where robustness (e.g., generalization across di stributions) is valued. Existing literature discussing this "hard-to-learn" conc ept are mainly expanded either along the dimension of the samples or the dimensi on of the features. In this paper, we aim to introduce a simple view merging the se two dimensions, leading to a new, simple yet effective, heuristic to train ma chine learning models by emphasizing the worst-cases on both the sample and the feature dimensions. We name our method W2D following the concept of "Worst-case along Two Dimensions". We validate the idea and demonstrate its empirical streng th over standard benchmarks.

****

Global Tracking Transformers

Xingyi Zhou, Tianwei Yin, Vladlen Koltun, Philipp Krähenbühl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp . 8771-8780

We present a novel transformer-based architecture for global multi-object tracki ng. Our network takes a short sequence of frames as input and produces global tr ajectories for all objects. The core component is a global tracking transformer that operates on objects from all frames in the sequence. The transformer encode s object features from all frames, and uses trajectory queries to group them int o trajectories. The trajectory queries are object features from a single frame a nd naturally produce unique trajectories. Our global tracking transformer does n ot require intermediate pairwise grouping or combinatorial association, and can be jointly trained with an object detector. It achieves competitive performance on the popular MOT17 benchmark, with 75.3 MOTA and 59.1 HOTA. More importantly, our framework seamlessly integrates into state-of-the-art large-vocabulary detec tors to track any objects. Experiments on the challenging TAO dataset show that our framework consistently improves upon baselines that are based on pairwise as sociation, outperforming published work by a significant 7.7 tracking mAP. Code is available at https://github.com/xingyizhou/GTR.

****

Backdoor Attacks on Self-Supervised Learning

Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Hamed Pirsiavash; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2022, pp. 13337-13346

Large-scale unlabeled data has spurred recent progress in self-supervised learni

ng methods that learn rich visual representations. State-of-the-art self-supervised methods for learning representations from images (e.g., MoCo, BYOL, MSF) use an inductive bias that random augmentations (e.g., random crops) of an image should produce similar embeddings. We show that such methods are vulnerable to backdoor attacks -- where an attacker poisons a small part of the unlabeled data by adding a trigger (image patch chosen by the attacker) to the images. The model performance is good on clean test images, but the attacker can manipulate the decision of the model by showing the trigger at test time. Backdoor attacks have been studied extensively in supervised learning and to the best of our knowledge, we are the first to study them for self-supervised learning. Backdoor attacks are more practical in self-supervised learning, since the use of large unlabeled data makes data inspection to remove poisons prohibitive. We show that in our targeted attack, the attacker can produce many false positives for the target category by using the trigger at test time. We also propose a defense method based on knowledge distillation that succeeds in neutralizing the attack. Our code is available here: https://github.com/UMBCvision/SSL-Backdoor

****

Multimodal Token Fusion for Vision Transformers

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12186-12195

Many adaptations of transformers have emerged to address the single-modal vision tasks, where self-attention modules are stacked to handle input sources like images. Intuitively, feeding multiple modalities of data to vision transformers could improve the performance, yet the inner-modal attentive weights may be diluted, which could thus greatly undermine the final performance. In this paper, we propose a multimodal token fusion method (TokenFusion), tailored for transformer-based vision tasks. To effectively fuse multiple modalities, TokenFusion dynamically detects uninformative tokens and substitute these tokens with projected and aggregated inter-modal features. Residual positional alignment is also adopted to enable explicit utilization of the inter-modal alignments after fusion. The design of TokenFusion allows the transformer to learn correlations among multimodal features, while the single-modal transformer architecture remains largely intact. Extensive experiments are conducted on a variety of homogeneous and heterogeneous modalities and demonstrate that TokenFusion surpasses state-of-the-art methods in three typical vision tasks: multimodal image-to-image translation, RGB-depth semantic segmentation, and 3D object detection with point cloud and images.

****

Exploring Frequency Adversarial Attacks for Face Forgery Detection

Shuai Jia, Chao Ma, Taiping Yao, Bangjie Yin, Shouhong Ding, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4103-4112

Various facial manipulation techniques have drawn serious public concerns in morality, security, and privacy. Although existing face forgery classifiers achieve promising performance on detecting fake images, these methods are vulnerable to adversarial examples with injected imperceptible perturbations on the pixels. Meanwhile, many face forgery detectors always utilize the frequency diversity between real and fake faces as a crucial clue. In this paper, instead of injecting adversarial perturbations into the spatial domain, we propose a frequency adversarial attack method against face forgery detectors. Concretely, we apply discrete cosine transform (DCT) on the input images and introduce a fusion module to capture the salient region of adversary in the frequency domain. Compared with existing adversarial attacks (e.g. FGSM, PGD) in the spatial domain, our method is more imperceptible to human observers and does not degrade the visual quality of the original images. Moreover, inspired by the idea of meta-learning, we also propose a hybrid adversarial attack that performs attacks in both the spatial and frequency domains. Extensive experiments indicate that the proposed method fools not only the spatial-based detectors but also the state-of-the-art frequency-based detectors effectively. In addition, the proposed frequency attack enhances

the transferability across face forgery detectors as black-box attacks.
****

GMFlow: Learning Optical Flow via Global Matching
Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8121-8130
Learning-based optical flow estimation has been dominated with the pipeline of cost volume with convolutions for flow regression, which is inherently limited to local correlations and thus is hard to address the long-standing challenge of large displacements. To alleviate this, the state-of-the-art framework RAFT gradually improves its prediction quality by using a large number of iterative refinements, achieving remarkable performance but introducing linearly increasing inference time. To enable both high accuracy and efficiency, we completely revamp the dominant flow regression pipeline by reformulating optical flow as a global matching problem, which identifies the correspondences by directly comparing feature similarities. Specifically, we propose a GMFlow framework, which consists of three main components: a customized Transformer for feature enhancement, a correlation and softmax layer for global feature matching, and a self-attention layer for flow propagation. We further introduce a refinement step that reuses GMFlow at higher feature resolution for residual flow prediction. Our new framework outperforms 31-refinements RAFT on the challenging Sintel benchmark, while using only one refinement and running faster, suggesting a new paradigm for accurate and efficient optical flow estimation. Code is available at https://github.com/haofeixu/gmflow.
****

Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation
Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10462-10472
Generating speech-consistent body and gesture movements is a long-standing problem in virtual avatar creation. Previous studies often synthesize pose movement in a holistic manner, where poses of all joints are generated simultaneously. Such a straightforward pipeline fails to generate fine-grained co-speech gestures. One observation is that the hierarchical semantics in speech and the hierarchical structures of human gestures can be naturally described into multiple granularities and associated together. To fully utilize the rich connections between speech audio and human gestures, we propose a novel framework named Hierarchical Audio-to-Gesture (HA2G) for co-speech gesture generation. In HA2G, a Hierarchical Audio Learner extracts audio representations across semantic granularities. A Hierarchical Pose Inferer subsequently renders the entire human pose gradually in a hierarchical manner. To enhance the quality of synthesized gestures, we develop a contrastive learning strategy based on audio-text alignment for better audio representations. Extensive experiments and human evaluation demonstrate that the proposed method renders realistic co-speech gestures and outperforms previous methods in a clear margin. Project page: https://alvinliu0.github.io/projects/HA2G.
****

FLAVA: A Foundational Language and Vision Alignment Model
Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, Douwe Kiela; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15638-15650
State-of-the-art vision and vision-and-language models rely on large-scale visio-linguistic pretraining for obtaining good performance on a variety of downstream tasks. Generally, such models are often either cross-modal (contrastive) or multi-modal (with earlier fusion) but not both; and they often only target specific modalities or tasks. A promising direction would be to use a single holistic universal model, as a "foundation", that targets all modalities at once---a true vision and language foundation model should be good at vision tasks, language tasks, and cross- and multi-modal vision and language tasks. We introduce FLAVA as such a model and demonstrate impressive performance on a wide range of 35 tasks

spanning these target modalities.
****
Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production
Ben Saunders, Necati Cihan Camgoz, Richard Bowden; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5141-5151
Sign languages are visual languages, with vocabularies as rich as their spoken language counterparts. However, current deep-learning based Sign Language Production (SLP) models produce under-articulated skeleton pose sequences from constrained vocabularies and this limits applicability. To be understandable and accepted by the deaf, an automatic SLP system must be able to generate co-articulated photo-realistic signing sequences for large domains of discourse. In this work, we tackle large-scale SLP by learning to co-articulate between dictionary signs, a method capable of producing smooth signing while scaling to unconstrained domains of discourse. To learn sign co-articulation, we propose a novel Frame Selection Network (FS-Net) that improves the temporal alignment of interpolated dictionary signs to continuous signing sequences. Additionally, we propose SignGAN, a pose-conditioned human synthesis model that produces photo-realistic sign language videos direct from skeleton pose. We propose a novel keypoint-based loss function which improves the quality of synthesized hand images. We evaluate our SLP model on the large-scale meineDGS (mDGS) corpus, conducting extensive user evaluation showing our FS-Net approach improves co-articulation of interpolated dictionary signs. Additionally, we show that SignGAN significantly outperforms all baseline methods for quantitative metrics, human perceptual studies and native deaf signer comprehension.
****
Explore Spatio-Temporal Aggregation for Insubstantial Object Detection: Benchmark Dataset and Baseline
Kailai Zhou, Yibo Wang, Tao Lv, Yunqian Li, Linsen Chen, Qiu Shen, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3104-3115
We endeavor on a rarely explored task named Insubstan-tial Object Detection (IOD), which aims to localize the object with following characteristics: (1) amorphous shape with indistinct boundary; (2) similarity to surroundings; (3) absence in color. Accordingly, it is far more challenging to distinguish insubstantial objects in a single static frame and the collaborative representation of spatial and tempo-ral information is crucial. Thus, we construct an IOD-Video dataset comprised of 600 videos (141,017 frames) covering various distances, sizes, visibility, and scenes captured by different spectral ranges. In addition, we develop a spatio-temporal aggregation framework for IOD, in which differ-ent backbones are deployed and a spatio-temporal aggregation loss (STAloss) is elaborately designed to leverage the consistency along the time axis. Experiments conducted on IOD-Video dataset demonstrate that spatio-temporal aggregation can significantly improve the performance of IOD. We hope our work will attract further researches into this valuable yet challenging task. The code will be available at: https://github.com/CalayZhou/IOD-Video.
****
OCSampler: Compressing Videos to One Clip With Single-Step Sampling
Jintao Lin, Haodong Duan, Kai Chen, Dahua Lin, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13894-13903
Videos incorporate rich semantics as well as redundant information. Seeking a compact yet effective video representation, e.g., sample informative frames from the entire video, is critical to efficient video recognition. There have been works that formulate frame sampling as a sequential decision task by selecting frames one by one according to their importance. In this paper, we present a more efficient framework named OCSampler, which explores such a representation with one short clip. OCSampler designs a new paradigm of learning instance-specific video condensation policies to select frames only in a single step. Rather than picking up frames sequentially like previous methods, we simply process a whole sequ

ence at once. Accordingly, these policies are derived from a light-weighted skim network together with a simple yet effective policy network. Moreover, we extend the proposed method with a frame number budget, enabling the framework to produce correct predictions in high confidence with as few frames as possible. Experiments on various benchmarks demonstrate the effectiveness of OCSampler over previous methods in terms of accuracy and efficiency. Specifically, it achieves 76.9% mAP and 21.7 GFLOPs on ActivityNet with an impressive throughput: 123.9 Video/s on a single TITAN Xp GPU.

****

Learning Bayesian Sparse Networks With Full Experience Replay for Continual Learning

Qingsen Yan, Dong Gong, Yuhang Liu, Anton van den Hengel, Javen Qinfeng Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 109-118

Continual Learning (CL) methods aim to enable machine learning models to learn new tasks without catastrophic forgetting of those that have been previously mastered. Existing CL approaches often keep a buffer of previously-seen samples, perform knowledge distillation, or use regularization techniques towards this goal. Despite their performance, they still suffer from interference across tasks which leads to catastrophic forgetting. To ameliorate this problem, we propose to only activate and select sparse neurons for learning current and past tasks at any stage. More parameters space and model capacity can thus be reserved for the future tasks. This minimizes the interference between parameters for different tasks. To do so, we propose a Sparse neural Network for Continual Learning (SNCL), which employs variational Bayesian sparsity priors on the activations of the neurons in all layers. Full Experience Replay (FER) provides effective supervision in learning the sparse activations of the neurons in different layers. A loss-aware reservoir-sampling strategy is developed to maintain the memory buffer. The proposed method is agnostic as to the network structures and the task boundaries. Experiments on different datasets show that SNCL achieves state-of-the-art result for mitigating forgetting.

****

Graph-Based Spatial Transformer With Memory Replay for Multi-Future Pedestrian Trajectory Prediction

Lihuan Li, Maurice Pagnucco, Yang Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2231-2241

Pedestrian trajectory prediction is an essential and challenging task for a variety of real-life applications such as autonomous driving and robotic motion planning. Besides generating a single future path, predicting multiple plausible future paths is becoming popular in some recent work on trajectory prediction. However, existing methods typically emphasize spatial interactions between pedestrians and surrounding areas but ignore the smoothness and temporal consistency of predictions. Our model aims to forecast multiple paths based on a historical trajectory by modeling multi-scale graph-based spatial transformers combined with a trajectory smoothing algorithm named "Memory Replay" utilizing a memory graph. Our method can comprehensively exploit the spatial information as well as correct the temporally inconsistent trajectories (e.g., sharp turns). We also propose a new evaluation metric named "Percentage of Trajectory Usage" to evaluate the comprehensiveness of diverse multi-future predictions. Our extensive experiments show that the proposed model achieves state-of-the-art performance on multi-future prediction and competitive results for single-future prediction. Code released at https://github.com/Jacobieee/ST-MR.

****

Scanline Homographies for Rolling-Shutter Plane Absolute Pose

Fang Bai, Agniva Sengupta, Adrien Bartoli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8993-9002

Cameras on portable devices are manufactured with a rolling-shutter (RS) mechanism, where the image rows (aka. scanlines) are read out sequentially. The unknown camera motions during the imaging process cause the so-called RS effects which are solved by motion assumptions in the literature. In this work, we give a solu

tion to the absolute pose problem free of motion assumptions. We categorically d
emonstrate that the only requirement is motion smoothness instead of stronger co
nstraints on the camera motion. To this end, we propose a novel mathematical abs
traction for RS cameras observing a planar scene, called the scanline-homography
, a 3x2 matrix with 5 DOFs. We establish the relationship between a scanline-hom
ography and the corresponding plane-homography, a 3x3 matrix with 6 DOFs assumin
g the camera is calibrated. We estimate the scanline-homographies of an RS frame
 using a smooth image warp powered by B-Splines, and recover the plane-homograph
ies afterwards to obtain the scanline-poses based on motion smoothness. We back
our claims with various experiments. Code and new datasets: https://bitbucket.or
g/clermontferrand/planarscanlinehomography/src/master/.
****

TableFormer: Table Structure Understanding With Transformers
Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, Peter Staar; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 4614-4623
Tables organize valuable content in a concise and compact representation. This c
ontent is extremely valuable for systems such as search engines, Knowledge Graph
's, etc, since they enhance their predictive capabilities. Unfortunately, tables
 come in a large variety of shapes and sizes. Furthermore, they can have complex
 column/row-header configurations, multiline rows, different variety of separati
on lines, missing entries, etc. As such, the correct identification of the table
-structure from an image is a non-trivial task. In this paper, we present a new
table-structure identification model. The latter improves the latest end-to-end
deep learning model (i.e. encoder-dual-decoder from PubTabNet) in two significan
t ways. First, we introduce a new object detection decoder for table-cells. In t
his way, we can obtain the content of the table-cells from programmatic PDF's di
rectly from the PDF source and avoid the training of the custom OCR decoders. Th
is architectural change leads to more accurate table-content extraction and allo
ws us to tackle non-english tables. Second, we replace the LSTM decoders with tr
ansformer based decoders. This upgrade improves significantly the previous state
-of-the-art tree-editing-distance-score (TEDS) from 91% to 98.5% on simple table
s and from 88.7% to 95% on complex tables.
****

Exemplar-Based Pattern Synthesis With Implicit Periodic Field Network
Haiwei Chen, Jiayi Liu, Weikai Chen, Shichen Liu, Yajie Zhao; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 3708-3717
Synthesis of ergodic, stationary visual patterns is widely applicable in texturi
ng, shape modeling, and digital content creation. The wide applicability of this
 technique thus requires the pattern synthesis approaches to be scalable, divers
e, and authentic. In this paper, we propose an exemplar-based visual pattern syn
thesis framework that aims to model the inner statistics of visual patterns and
generate new, versatile patterns that meet the aforementioned requirements. To t
his end, we propose an implicit network based on generative adversarial network
(GAN) and periodic encoding, thus calling our network the Implicit Periodic Fiel
d Network (IPFN). The design of IPFN ensures scalability: the implicit formulati
on directly maps the input coordinates to features, which enables synthesis of a
rbitrary size and is computationally efficient for 3D shape synthesis. Learning
with a periodic encoding scheme encourages diversity: the network is constrained
 to model the inner statistics of the exemplar based on spatial latent codes in
a periodic field. Coupled with continuously designed GAN training procedures, IP
FN is shown to synthesize tileable patterns with smooth transitions and local va
riations. Last but not least, thanks to both the adversarial training technique
and the encoded Fourier features, IPFN learns high-frequency functions that prod
uce authentic, high-quality results. To validate our approach, we present novel
experimental results on various applications in 2D texture synthesis and 3D shap
e synthesis.
****

Grounded Language-Image Pre-Training

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yi wu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, Jianf eng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 10965-10975

This paper presents a grounded language-image pre-training (GLIP) model for lear ning object-level, language-aware, and semantic-rich visual representations. GLI P unifies object detection and phrase grounding for pre-training. The unificatio n brings two benefits: 1) it allows GLIP to learn from both detection and ground ing data to improve both tasks and bootstrap a good grounding model; 2) GLIP can leverage massive image-text pairs by generating grounding boxes in a self-train ing fashion, making the learned representations semantic-rich. In our experiment s, we pre-train GLIP on 27M grounding data, including 3M human-annotated and 24M web-crawled image-text pairs. The learned representations demonstrate strong ze ro-shot and few-shot transferability to various object-level recognition tasks. 1) When directly evaluated on COCO and LVIS (without seeing any images in COCO d uring pre-training), GLIP achieves 49.8 AP and 26.9 AP, respectively, surpassing many supervised baselines. 2) After fine-tuned on COCO, GLIP achieves 60.8 AP o n val and 61.5 AP on test-dev, surpassing prior SoTA. 3) When transferred to 13 downstream object detection tasks, a 1-shot GLIP rivals with a fully-supervised Dynamic Head.

****

Spectral Unsupervised Domain Adaptation for Visual Recognition

Jingyi Zhang, Jiaxing Huang, Zichen Tian, Shijian Lu; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9829-9 840

Though unsupervised domain adaptation (UDA) has achieved very impressive progres s recently, it remains a great challenge due to missing target annotations and t he rich discrepancy between source and target distributions. We propose Spectral UDA (SUDA), an effective and efficient UDA technique that works in the spectral space and can generalize across different visual recognition tasks. SUDA addres ses the UDA challenges from two perspectives. First, it introduces a spectrum tr ansformer (ST) that mitigates inter-domain discrepancies by enhancing domain-inv ariant spectra while suppressing domain-variant spectra of source and target sam ples simultaneously. Second, it introduces multi-view spectral learning that lea rns useful unsupervised representations by maximizing mutual information among m ultiple ST-generated spectral views of each target sample. Extensive experiments show that SUDA achieves superior accuracy consistently across different visual tasks in image classification, semantic segmentation, and object detection. Addi tionally, SUDA also works with the transformer-based network and achieves state- of-the-art performance on object detection.

****

AdaInt: Learning Adaptive Intervals for 3D Lookup Tables on Real-Time Image Enha ncement

Canqian Yang, Meiguang Jin, Xu Jia, Yi Xu, Ying Chen; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17522- 17531

The 3D Lookup Table (3D LUT) is a highly-efficient tool for real-time image enha ncement tasks, which models a non-linear 3D color transform by sparsely sampling it into a discretized 3D lattice. Previous works have made efforts to learn ima ge-adaptive output color values of LUTs for flexible enhancement but neglect the importance of sampling strategy. They adopt a sub-optimal uniform sampling poin t allocation, limiting the expressiveness of the learned LUTs since the (tri-)li near interpolation between uniform sampling points in the LUT transform might fa il to model local non-linearities of the color transform. Focusing on this probl em, we present AdaInt (Adaptive Intervals Learning), a novel mechanism to achiev e a more flexible sampling point allocation by adaptively learning the non-unifo rm sampling intervals in the 3D color space. In this way, a 3D LUT can increase its capability by conducting dense sampling in color ranges requiring highly non -linear transforms and sparse sampling for near-linear transforms. The proposed AdaInt could be implemented as a compact and efficient plug-and-play module for

a 3D LUT-based method. To enable the end-to-end learning of AdaInt, we design a novel differentiable operator called AiLUT-Transform (Adaptive Interval LUT Transform) to locate input colors in the non-uniform 3D LUT and provide gradients to the sampling intervals. Experiments demonstrate that methods equipped with AdaInt can achieve state-of-the-art performance on two public benchmark datasets with a negligible overhead increase. Our source code is available at https://github.com/ImCharlesY/AdaInt.
****

PatchFormer: An Efficient Point Transformer With Patch Attention
Cheng Zhang, Haocheng Wan, Xinyi Shen, Zizhao Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11799-11808
The point cloud learning community is witnesses a modeling shift from CNNs to Transformers, where pure Transformer architectures have achieved top accuracy on the major learning benchmarks. However, existing point Transformers are computationally expensive since they need to generate a large attention map, which has quadratic complexity (both in space and time) with respect to input size. To solve this shortcoming, we introduce patch-attention (PAT) to adaptively learn a much smaller set of bases upon which the attention maps are computed. By a weighted summation upon these bases, PAT not only captures the global shape context but also achieves linear complexity to input size. In addition, we propose a lightweight Multi-Scale Attention (MST) block to build attentions among features of different scales, providing the model with multi-scale features. Equipped with the PAT and MST, we construct our neural architecture called PatchFormer that integrates both modules into a joint framework for point cloud learning. Extensive experiments demonstrate that our network achieves comparable accuracy on general point cloud learning tasks with 9.2x speed-up than previous point Transformers.
****

Recurrent Glimpse-Based Decoder for Detection With Transformer
Zhe Chen, Jing Zhang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5260-5269
Although detection with Transformer (DETR) is increasingly popular, its global attention modeling requires an extremely long training period to optimize and achieve promising detection performance. Alternative to existing studies that mainly develop advanced feature or embedding designs to tackle the training issue, we point out that the Region-of-Interest (RoI) based detection refinement can easily help mitigate the difficulty of training for DETR methods. Based on this, we introduce a novel REcurrent Glimpse-based decOder (REGO) in this paper. In particular, the REGO employs a multi-stage recurrent processing structure to help the attention of DETR gradually focus on foreground objects more accurately. In each processing stage, visual features are extracted as glimpse features from RoIs with enlarged bounding box areas of detection results from the previous stage. Then, a glimpse-based decoder is introduced to provide refined detection results based on both the glimpse features and the attention modeling outputs of the previous stage. In practice, REGO can be easily embedded in representative DETR variants while maintaining their fully end-to-end training and inference pipelines. In particular, REGO helps Deformable DETR achieve 44.8 AP on the MSCOCO dataset with only 36 training epochs, compared with the first DETR and the Deformable DETR that require 500 and 50 epochs to achieve comparable performance, respectively. Experiments also show that REGO consistently boosts the performance of different DETR detectors by up to 7% relative gain at the same setting of 50 training epochs. Code is available via https://github.com/zhechen/Deformable-DETR-REGO.
****

Generating 3D Bio-Printable Patches Using Wound Segmentation and Reconstruction To Treat Diabetic Foot Ulcers
Han Joo Chae, Seunghwan Lee, Hyewon Son, Seungyeob Han, Taebin Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2539-2549
We introduce AiD Regen, a novel system that generates 3D wound models combining 2D semantic segmentation with 3D reconstruction so that they can be printed via

3D bio-printers during the surgery to treat diabetic foot ulcers (DFUs). AiD Regen seamlessly binds the full pipeline, which includes RGB-D image capturing, semantic segmentation, boundary-guided point-cloud processing, 3D model reconstruction, and 3D printable G-code generation, into a single system that can be used out of the box. We developed a multi-stage data preprocessing method to handle small and unbalanced DFU image datasets. AiD Regen's human-in-the-loop machine learning interface enables clinicians to not only create 3D regenerative patches with just a few touch interactions but also customize and confirm wound boundaries. As evidenced by our experiments, our model outperforms prior wound segmentation models and our reconstruction algorithm is capable of generating 3D wound models with compelling accuracy. We further conducted a case study on a real DFU patient and demonstrated the effectiveness of AiD Regen in treating DFU wounds.

****

SimMIM: A Simple Framework for Masked Image Modeling

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, Han Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9653-9663

This paper presents SimMIM, a simple framework for masked image modeling. We have simplified recently proposed relevant approaches, without the need for special designs, such as block-wise masking and tokenization via discrete VAE or clustering. To investigate what makes a masked image modeling task learn good representations, we systematically study the major components in our framework, and find that the simple designs of each component have revealed very strong representation learning performance: 1) random masking of the input image with a moderately large masked patch size (e.g., 32) makes a powerful pre-text task; 2) predicting RGB values of raw pixels by direct regression performs no worse than the patch classification approaches with complex designs; 3) the prediction head can be as light as a linear layer, with no worse performance than heavier ones. Using ViT-B, our approach achieves 83.8% top-1 fine-tuning accuracy on ImageNet-1K by pre-training also on this dataset, surpassing previous best approach by +0.6%. When applied to a larger model with about 650 million parameters, SwinV2-H, it achieves 87.1% top-1 accuracy on ImageNet-1K using only ImageNet-1K data. We also leverage this approach to address the data-hungry issue faced by large-scale model training, that a 3B model (SwinV2-G) is successfully trained to achieve state-of-the-art accuracy on four representative vision benchmarks using 40x less labeled data than that in previous practice (JFT-3B). The code is available at https://github.com/microsoft/SimMIM.

****

OmniFusion: 360 Monocular Depth Estimation via Geometry-Aware Fusion

Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, Liu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2801-2810

A well-known challenge in applying deep-learning methods to omnidirectional images is spherical distortion. In dense regression tasks such as depth estimation, where structural details are required, using a vanilla CNN layer on the distorted 360 image results in undesired information loss. In this paper, we propose a 360 monocular depth estimation pipeline, OmniFusion, to tackle the spherical distortion issue. Our pipeline transforms a 360 image into less-distorted perspective patches (i.e. tangent images) to obtain patch-wise predictions via CNN, and then merge the patch-wise results for final output. To handle the discrepancy between patch-wise predictions which is a major issue affecting the merging quality, we propose a new framework with the following key components. First, we propose a geometry-aware feature fusion mechanism that combines 3D geometric features with 2D image features to compensate for the patch-wise discrepancy. Second, we employ the self-attention-based transformer architecture to conduct a global aggregation of patch-wise information, which further improves the consistency. Last, we introduce an iterative depth refinement mechanism, to further refine the estimated depth based on the more accurate geometric features. Experiments show that our method greatly mitigates the distortion issue, and achieves state-of-the-art performances on several 360 monocular depth estimation benchmark datasets.

****
Label Matching Semi-Supervised Object Detection
Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu , Mingli Song, Yueting Zhuang; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 14381-14390
Semi-supervised object detection has made significant progress with the developm ent of mean teacher driven self-training. Despite the promising results, the lab el mismatch problem is not yet fully explored in the previous works, leading to severe confirmation bias during self-training. In this paper, we delve into this problem and propose a simple yet effective LabelMatch framework from two differ ent yet complementary perspectives, i.e., distribution-level and instance-level. For the former one, it is reasonable to approximate the class distribution of t he unlabeled data from that of the labeled data according to Monte Carlo Samplin g. Guided by this weakly supervision cue, we introduce a re-distribution mean te acher, which leverages adaptive label-distribution-aware confidence thresholds t o generate unbiased pseudo labels to drive student learning. For the latter one, there exists an overlooked label assignment ambiguity problem across teacher-st udent models. To remedy this issue, we present a novel label assignment mechanis m for self-training framework, namely proposal self-assignment, which injects th e proposals from student into teacher and generates accurate pseudo labels to ma tch each proposal in the student model accordingly. Experiments on both MS-COCO and PASCAL-VOC datasets demonstrate the considerable superiority of our proposed framework to other state-of-the-arts. Code will be available at https://github. com/HIK-LAB/SSOD.
****
RegionCLIP: Region-Based Language-Image Pretraining
Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Ha rold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16793-16803
Contrastive language-image pretraining (CLIP) using image-text pairs has achieve d impressive results on image classification in both zero-shot and transfer lear ning settings. However, we show that directly applying such models to recognize image regions for object detection leads to unsatisfactory performance due to a major domain shift: CLIP was trained to match an image as a whole to a text desc ription, without capturing the fine-grained alignment between image regions and text spans. To mitigate this issue, we propose a new method called RegionCLIP th at significantly extends CLIP to learn region-level visual representations, thus enabling fine-grained alignment between image regions and textual concepts. Our method leverages a CLIP model to match image regions with template captions, an d then pretrains our model to align these region-text pairs in the feature space . When transferring our pretrained model to the open-vocabulary object detection task, our method outperforms the state of the art by 3.8 AP50 and 2.2 AP for no vel categories on COCO and LVIS datasets, respectively. Further, the learned reg ion representations support zero-shot inference for object detection, showing pr omising results on both COCO and LVIS datasets. Our code is available at https:/ /github.com/microsoft/RegionCLIP.
****
Video Frame Interpolation Transformer
Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17482-17491
Existing methods for video interpolation heavily rely on deep convolution neural networks, and thus suffer from their intrinsic limitations, such as content-agn ostic kernel weights and restricted receptive field. To address these issues, we propose a Transformer-based video interpolation framework that allows content-a ware aggregation weights and considers long-range dependencies with the self-att ention operations. To avoid the high computational cost of global self-attention , we introduce the concept of local attention into video interpolation and exten d it to the spatial-temporal domain. Furthermore, we propose a space-time separa

tion strategy to save memory usage, which also improves performance. In addition , we develop a multi-scale frame synthesis scheme to fully realize the potential of Transformers. Extensive experiments demonstrate the proposed model performs favorably against the state-of-the-art methods both quantitatively and qualitati vely on a variety of benchmark datasets. The code and models are released at htt ps://github.com/zhshi0816/ Video-Frame-Interpolation-Transformer.
****

An MIL-Derived Transformer for Weakly Supervised Point Cloud Segmentation
Cheng-Kun Yang, Ji-Jia Wu, Kai-Syun Chen, Yung-Yu Chuang, Yen-Yu Lin; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11830-11839
We address weakly supervised point cloud segmentation by proposing a new model, MIL-derived transformer, to mine additional supervisory signals. First, the tran sformer model is derived based on multiple instance learning (MIL) to explore pa ir-wise cloud-level supervision, where two clouds of the same category yield a p ositive bag while two of different classes produce a negative bag. It leverages not only individual cloud annotations but also pair-wise cloud semantics for mod el optimization. Second, Adaptive global weighted pooling (AdaGWP) is integrated into our transformer model to replace max pooling and average pooling. It intro duces learnable weights to re-scale logits in the class activation maps. It is m ore robust to noise while discovering more complete foreground points under weak supervision. Third, we perform point subsampling and enforce feature equivarian ce between the original and subsampled point clouds for regularization. The prop osed method is end-to-end trainable and is general because it can work with diff erent backbones with diverse types of weak supervision signals, including sparse ly annotated points and cloud-level labels. The experiments show that it achieve s state-of-the-art performance on the S3DIS and ScanNet benchmarks.
****

Fast Light-Weight Near-Field Photometric Stereo
Daniel Lichy, Soumyadip Sengupta, David W. Jacobs; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12612-126 21
We introduce the first end-to-end learning-based solution to near-field Photomet ric Stereo (PS), where the light sources are close to the object of interest. Th is setup is especially useful for reconstructing large immobile objects. Our met hod is fast, producing a mesh from 52 512x384 resolution images in about 1 secon d on a commodity GPU, thus potentially unlocking several AR/VR applications. Exi sting approaches rely on optimization coupled with a far-field PS network operat ing on pixels or small patches. Using optimization makes these approaches slow a nd memory intensive (requiring 17GB GPU and 27GB of CPU memory) while using only pixels or patches makes them highly susceptible to noise and calibration errors . To address these issues, we develop a recursive multi-resolution scheme to est imate surface normal and depth maps of the whole image at each step. The predict ed depth map at each scale is then used to estimate 'per-pixel lighting' for the next scale. This design makes our approach almost 45x faster and 2 degrees more accurate (11.3 vs. 13.3 degrees Mean Angular Error) than the state-of-the-art n ear-field PS reconstruction technique, which uses iterative optimization.
****

BCOT: A Markerless High-Precision 3D Object Tracking Benchmark
Jiachen Li, Bin Wang, Shiqiang Zhu, Xin Cao, Fan Zhong, Wenxuan Chen, Te Li, Jas on Gu, Xueying Qin; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 6697-6706
Template-based 3D object tracking still lacks a high-precision benchmark of real scenes due to the difficulty of annotating the accurate 3D poses of real moving video objects without using markers. In this paper, we present a multi-view app roach to estimate the accurate 3D poses of real moving objects, and then use bin ocular data to construct a new benchmark for monocular textureless 3D object tra cking. The proposed method requires no markers, and the cameras only need to be synchronous, relatively fixed as cross-view and calibrated. Based on our object-centered model, we jointly optimize the object pose by minimizing shape re-proje

ction constraints in all views, which greatly improves the accuracy compared wit h the single-view approach, and is even more accurate than the depth-based metho d. Our new benchmark dataset contains 20 textureless objects, 22 scenes, 404 vid eo sequences and 126K images captured in real scenes. The annotation error is gu aranteed to be less than 2mm, according to both theoretical analysis and validat ion experiments. We re-evaluate the state-of-the-art 3D object tracking methods with our dataset, reporting their performance ranking in real scenes. Our BCOT b enchmark and code can be found at https://ar3dv.github.io/BCOT-Benchmark/.
****

Omni-DETR: Omni-Supervised Object Detection With Transformers
Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2022, pp. 9367-9376
We consider the problem of omni-supervised object detection, which can use unlab eled, fully labeled and weakly labeled annotations, such as image tags, counts, points, etc., for object detection. This is enabled by a unified architecture, O mni-DETR, based on the recent progress on student-teacher framework and end-to-e nd transformer based object detection. Under this unified architecture, differen t types of weak labels can be leveraged to generate accurate pseudo labels, by a bipartite matching based filtering mechanism, for the model to learn. In the ex periments, Omni-DETR has achieved state-of-the-art results on multiple datasets and settings. And we have found that weak annotations can help to improve detect ion performance and a mixture of them can achieve a better trade-off between ann otation cost and accuracy than the standard complete annotation. These findings could encourage larger object detection datasets with mixture annotations. The c ode is available at https://github.com/amazon-research/omni-detr.
****

Uniform Subdivision of Omnidirectional Camera Space for Efficient Spherical Ster eo Matching
Donghun Kang, Hyeonjoong Jang, Jungeon Lee, Chong-Min Kyung, Min H. Kim; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12972-12980
Omnidirectional cameras have been used widely to better understand surrounding e nvironments. They are often configured as stereo to estimate depth. However, due to the optics of the fisheye lens, conventional epipolar geometry is inapplicab le directly to omnidirectional camera images. Intermediate formats of omnidirect ional images, such as equirectangular images, have been used. However, stereo ma tching performance on these image formats has been lower than the conventional s tereo due to severe image distortion near pole regions. In this paper, to addres s the distortion problem of omnidirectional images, we devise a novel subdivisio n scheme of a spherical geodesic grid. This enables more isotropic patch samplin g of spherical image information in the omnidirectional camera space. Our spheri cal geodesic grid is tessellated with an equal-arc subdivision, making the cell sizes and in-between distances as uniform as possible, i.e., the arc length of t he spherical grid cell's edges is well regularized. Also, our uniformly tessella ted coordinates in a 2D image can be transformed into spherical coordinates via one-to-one mapping, allowing for analytical forward/backward transformation. Our uniform tessellation scheme achieves a higher accuracy of stereo matching than the traditional cylindrical and cubemap-based approaches, reducing the memory fo otage required for stereo matching by 20 %.
****

High-Resolution Image Synthesis With Latent Diffusion Models
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684-10695
By decomposing the image formation process into a sequential application of deno ising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis re sults on image data and beyond. Additionally, their formulation allows for a gui ding mechanism to control the image generation process without retraining. Howev er, since these models typically operate directly in pixel space, optimization o

f powerful DMs often consumes hundreds of GPU days and inference is expensive du e to sequential evaluations. To enable DM training on limited computational reso urces while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, traini ng diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, great ly boosting visual fidelity. By introducing cross-attention layers into the mode l architecture, we turn diffusion models into powerful and flexible generators f or general conditioning inputs such as text or bounding boxes and high-resolutio n synthesis becomes possible in a convolutional manner. Our latent diffusion mod els (LDMs) achieve new state of the art scores for image inpainting and class-co nditional image synthesis and highly competitive performance on various tasks, i ncluding unconditional image generation, text-to-image synthesis, and super-reso lution, while significantly reducing computational requirements compared to pixe l-based DMs.

****

Improving Adversarially Robust Few-Shot Image Classification With Generalizable Representations

Junhao Dong, Yuan Wang, Jian-Huang Lai, Xiaohua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9025-9034

Few-Shot Image Classification (FSIC) aims to recognize novel image classes with limited data, which is significant in practice. In this paper, we consider the F SIC problem in the case of adversarial examples. This is an extremely challengin g issue because current deep learning methods are still vulnerable when handling adversarial examples, even with massive labeled training samples. For this prob lem, existing works focus on training a network in the meta-learning fashion tha t depends on numerous sampled few-shot tasks. In comparison, we propose a simple but effective baseline through directly learning generalizable representations without tedious task sampling, which is robust to unforeseen adversarial FSIC ta sks. Specifically, we introduce an adversarial-aware mechanism to establish auxi liary supervision via feature-level differences between legitimate and adversari al examples. Furthermore, we design a novel adversarial-reweighted training mann er to alleviate the imbalance among adversarial examples. The feature purifier i s also employed as post-processing for adversarial features. Moreover, our metho d can obtain generalizable representations to remain superior transferability, e ven facing cross-domain adversarial examples. Extensive experiments show that ou r method can significantly outperform state-of-the-art adversarially robust FSIC methods on two standard benchmarks.

****

Transferable Sparse Adversarial Attack

Ziwen He, Wei Wang, Jing Dong, Tieniu Tan; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14963-14972

Deep neural networks have shown their vulnerability to adversarial attacks. In t his paper, we focus on sparse adversarial attack based on the l_0 norm constrain t, which can succeed by only modifying a few pixels of an image. Despite a high attack success rate, prior sparse attack methods achieve a low transferability u nder the black-box protocol due to overfitting the target model. Therefore, we i ntroduce a generator architecture to alleviate the overfitting issue and thus ef ficiently craft transferable sparse adversarial examples. Specifically, the gene rator decouples the sparse perturbation into amplitude and position components. We carefully design a random quantization operator to optimize these two compone nts jointly in an end-to-end way. The experiment shows that our method has impro ved the transferability by a large margin under a similar sparsity setting compa red with state-of-the-art methods. Moreover, our method achieves superior infere nce speed, 700 times faster than other optimization-based methods. The code is a vailable at https://github.com/shaguopohuaizhe/TSAA.

****

CREAM: Weakly Supervised Object Localization via Class RE-Activation Mapping

Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan L

u, Shang Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2022, pp. 9437-9446
Weakly Supervised Object Localization (WSOL) aims to localize objects with image -level supervision. Existing works mainly rely on Class Activation Mapping (CAM) derived from a classification model. However, CAM-based methods usually focus o n the most discriminative parts of an object (i.e., incomplete localization prob lem). In this paper, we empirically prove that this problem is associated with t he mixup of the activation values between less discriminative foreground regions and the background. To address it, we propose Class RE-Activation Mapping (CREA M), a novel clustering-based approach to boost the activation values of the inte gral object regions. To this end, we introduce class-specific foreground and bac kground context embeddings as cluster centroids. A CAM-guided momentum preservat ion strategy is developed to learn the context embeddings during training. At th e inference stage, the re-activation mapping is formulated as a parameter estima tion problem under Gaussian Mixture Model, which can be solved by deriving an un supervised Expectation-Maximization based soft-clustering algorithm. By simply i ntegrating CREAM into various WSOL approaches, our method significantly improves their performance. CREAM achieves the state-of-the-art performance on CUB, ILSV RC and OpenImages benchmark datasets. Code is available at https://github.com/JL Repo/CREAM.
****
Semi-Weakly-Supervised Learning of Complex Actions From Instructional Task Video s
Yuhan Shen, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3344-3354
We address the problem of action segmentation in instructional task videos with a small number of weakly-labeled training videos and a large number of unlabeled videos, which we refer to as Semi-Weakly-Supervised Learning (SWSL) of actions. We propose a general SWSL framework that can efficiently learn from both types of videos and can leverage any of the existing weakly-supervised action segmenta tion methods. Our key observation is that the distance between the transcript of an unlabeled video and those of the weakly-labeled videos from the same task is small yet often nonzero. Therefore, we develop a Soft Restricted Edit (SRE) los s to encourage small variations between the predicted transcripts of unlabeled v ideos and ground-truth transcripts of the weakly-labeled videos of the same task . To compute the SRE loss, we develop a flexible transcript prediction (FTP) met hod that uses the output of the action classifier to find both the length of the transcript and the sequence of actions occurring in an unlabeled video. We prop ose an efficient learning scheme in which we alternate between minimizing our pr oposed loss and generating pseudo-transcripts for unlabeled videos. By experimen ts on two benchmark datasets, we demonstrate that our approach can significantly improve the performance by using unlabeled videos, especially when the number o f weakly-labeled videos is small.
****
APRIL: Finding the Achilles' Heel on Privacy for Vision Transformers
Jiahao Lu, Xi Sheryl Zhang, Tianli Zhao, Xiangyu He, Jian Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10051-10060
Federated learning frameworks typically require collaborators to share their loc al gradient updates of a common model instead of sharing training data to preser ve privacy. However, prior works on Gradient Leakage Attacks showed that private training data can be revealed from gradients. So far almost all relevant works base their attacks on fully-connected or convolutional neural networks. Given th e recent overwhelmingly rising trend of adapting Transformers to solve multifari ous vision tasks, it is highly important to investigate the privacy risk of visi on transformers. In this paper, we analyse the gradient leakage risk of self-att ention based mechanism in both theoretical and practical manners. Particularly, we propose APRIL - Attention PRIvacy Leakage, which poses a strong threat to sel f-attention inspired models such as ViT. Showing how vision Transformers are at the risk of privacy leakage via gradients, we urge the significance of designing

privacy-safer Transformer models and defending schemes.
****
Text Spotting Transformers
Xiang Zhang, Yongwen Su, Subarna Tripathi, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9519-9528
In this paper, we present TExt Spotting TRansformers (TESTR), a generic end-to-end text spotting framework using Transformers for text detection and recognition in the wild. TESTR builds upon a single encoder and dual decoders for the joint text-box control point regression and character recognition. Other than most existing literature, our method is free from Region-of-Interest operations and heuristics-driven post-processing procedures; TESTR is particularly effective when dealing with curved text-boxes where special cares are needed for the adaptation of the traditional bounding-box representations. We show our canonical representation of control points suitable for text instances in both Bezier curve and polygon annotations. In addition, we design a bounding-box guided polygon detection (box-to-polygon) process. Experiments on curved and arbitrarily shaped datasets demonstrate state-of-the-art performances of the proposed TESTR algorithm.
****
Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields
Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, Peter Hedman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5470-5479
Though neural radiance fields ("NeRF") have demonstrated impressive view synthesis results on objects and small bounded regions of space, they struggle on "unbounded" scenes, where the camera may point in any direction and content may exist at any distance. In this setting, existing NeRF-like models often produce blurry or low-resolution renderings (due to the unbalanced detail and scale of nearby and distant objects), are slow to train, and may exhibit artifacts due to the inherent ambiguity of the task of reconstructing a large scene from a small set of images. We present an extension of mip-NeRF (a NeRF variant that addresses sampling and aliasing) that uses a non-linear scene parameterization, online distillation, and a novel distortion-based regularizer to overcome the challenges presented by unbounded scenes. Our model, which we dub "mip-NeRF 360" as we target scenes in which the camera rotates 360 degrees around a point, reduces mean-squared error by 57% compared to mip-NeRF, and is able to produce realistic synthesized views and detailed depth maps for highly intricate, unbounded real-world scenes.
****
VALHALLA: Visual Hallucination for Machine Translation
Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu (Richard) Chen, Rogerio S. Feris, David Cox, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5216-5226
Designing better machine translation systems by considering auxiliary inputs such as images has attracted much attention in recent years. While existing methods show promising performance over the conventional text-only translation systems, they typically require paired text and image as input during inference, which limits their applicability to real-world scenarios. In this paper, we introduce a visual hallucination framework, called VALHALLA, which requires only source sentences at inference time and instead uses hallucinated visual representations for multimodal machine translation. In particular, given a source sentence an autoregressive hallucination transformer is used to predict a discrete visual representation from the input text, and the combined text and hallucinated representations are utilized to obtain the target translation. We train the hallucination transformer jointly with the translation transformer using standard backpropagation with cross-entropy losses while being guided by an additional loss that encourages consistency between predictions using either ground-truth or hallucinated visual representations. Extensive experiments on three standard translation data sets with a diverse set of language pairs demonstrate the effectiveness of our approach over both text-only baselines and state-of-the-art methods. Our codes an

d models will be publicly available. Project page: http://www.svcl.ucsd.edu/projects/valhalla.
****

StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation
Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13503-13513
We introduce a high resolution, 3D-consistent image and shape generation technique which we call StyleSDF. Our method is trained on single view RGB data only, and stands on the shoulders of StyleGAN2 for image generation, while solving two main challenges in 3D-aware GANs: 1) high-resolution, view-consistent generation of the RGB images, and 2) detailed 3D shape. We achieve this by merging an SDF-based 3D representation with a style-based 2D generator. Our 3D implicit network renders low-resolution feature maps, from which the style-based network generates view-consistent, 1024x1024 images. Notably, our SDF-based 3D modeling defines detailed 3D surfaces, leading to consistent volume rendering. Our method shows higher quality results compared to state of the art in terms of visual and geometric quality.
****

Incorporating Semi-Supervised and Positive-Unlabeled Learning for Boosting Full Reference Image Quality Assessment
Yue Cao, Zhaolin Wan, Dongwei Ren, Zifei Yan, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5851-5861
Full-reference (FR) image quality assessment (IQA) evaluates the visual quality of a distorted image by measuring its perceptual difference with pristine-quality reference, and has been widely used in low level vision tasks. Pairwise labeled data with mean opinion score (MOS) are required in training FR-IQA model, but is time-consuming and cumbersome to collect. In contrast, unlabeled data can be easily collected from an image degradation or restoration process, making it encouraging to exploit unlabeled training data to boost FR-IQA performance. Moreover, due to the distribution inconsistency between labeled and unlabeled data, outliers may occur in unlabeled data, further increasing the training difficulty. In this paper, we suggest to incorporate semi-supervised and positive-unlabeled (PU) learning for exploiting unlabeled data while mitigating the adverse effect of outliers. Particularly, by treating all labeled data as positive samples, PU learning is leveraged to identify negative samples (i.e., outliers) from unlabeled data. Semi-supervised learning (SSL) is further deployed to exploit positive unlabeled data by dynamically generating pseudo-MOS. We adopt a dual-branch network including reference and distortion branches. Furthermore, spatial attention is introduced in the reference branch to concentrate more on the informative regions, and sliced Wasserstein distance is used for robust difference map computation to address the misalignment issues caused by images recovered by GAN models. Extensive experiments show that our method performs favorably against state-of-the-arts on the benchmark datasets PIPAL, KADID-10k, TID2013, LIVE and CSIQ.
****

GLAMR: Global Occlusion-Aware Human Mesh Recovery With Dynamic Cameras
Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11038-11049
We present an approach for 3D global human mesh recovery from monocular videos recorded with dynamic cameras. Our approach is robust to severe and long-term occlusions and tracks human bodies even when they go outside the camera's field of view. To achieve this, we first propose a deep generative motion infiller, which autoregressively infills the body motions of occluded humans based on visible motions. Additionally, in contrast to prior work, our approach reconstructs human meshes in consistent global coordinates even with dynamic cameras. Since the joint reconstruction of human motions and camera poses is underconstrained, we propose a global trajectory predictor that generates global human trajectories based on local body movements. Using the predicted trajectories as anchors, we prese

nt a global optimization framework that refines the predicted trajectories and o
ptimizes the camera poses to match the video evidence such as 2D keypoints. Expe
riments on challenging indoor and in-the-wild datasets with dynamic cameras demo
nstrate that the proposed approach outperforms prior methods significantly in te
rms of motion infilling and global mesh recovery.
****

HINT: Hierarchical Neuron Concept Explainer
Andong Wang, Wei-Ning Lee, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10254-10264
To interpret deep networks, one main approach is to associate neurons with human
-understandable concepts. However, existing methods often ignore the inherent co
nnections of different concepts (e.g., dog and cat both belong to animals), and
thus lose the chance to explain neurons responsible for higher-level concepts (e
.g., animal). In this paper, we study hierarchical concepts inspired by the hier
archical cognition process of human beings. To this end, we propose HIerarchical
 Neuron concepT explainer (HINT) to effectively build bidirectional associations
 between neurons and hierarchical concepts in a low-cost and scalable manner. HI
NT enables us to systematically and quantitatively study whether and how the imp
licit hierarchical relationships of concepts are embedded into neurons. Specific
ally, HINT identifies collaborative neurons responsible for one concept and mult
imodal neurons pertinent to different concepts, at different semantic levels fro
m concrete concepts (e.g., dog) to more abstract ones (e.g., animal). Finally, w
e verify the faithfulness of the associations using Weakly Supervised Object Loc
alization, and demonstrate its applicability in various tasks, such as discoveri
ng saliency regions and explaining adversarial attacks. Code is available on htt
ps://github.com/AntonotnaWang/HINT.
****

Capturing and Inferring Dense Full-Body Human-Scene Contact
Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Ale
xiadis, Senya Polikovsky, Daniel Scharstein, Michael J. Black; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p
p. 13274-13285
Inferring human-scene contact (HSC) is the first step toward understanding how h
umans interact with their surroundings. While detecting 2D human-object interact
ion (HOI) and reconstructing 3D human pose and shape (HPS) have enjoyed signific
ant progress, reasoning about 3D human-scene contact from a single image is stil
l challenging. Existing HSC detection methods consider only a few types of prede
fined contact, often reduce body and scene to a small number of primitives, and
even overlook image evidence. To predict human-scene contact from a single image
, we address the limitations above from both data and algorithmic perspectives.
We capture a new dataset called RICH for "Real scenes, Interaction, Contact and
Humans." RICH contains multiview outdoor/indoor video sequences at 4K resolution
, ground-truth 3D human bodies captured using markerless motion capture, 3D body
 scans, and high resolution 3D scene scans. A key feature of RICH is that it als
o contains accurate vertex-level contact labels on the body. Using RICH, we trai
n a network that predicts dense body-scene contacts from a single RGB image. Our
 key insight is that regions in contact are always occluded so the network needs
 the ability to explore the whole image for evidence. We use a transformer to le
arn such non-local relationships and propose a new Body-Scene contact TRansfOrme
r (BSTRO). Very few methods explore 3D contact; those that do focus on the feet
only, detect foot contact as a post-processing step, or infer contact from body
pose without looking at the scene. To our knowledge, BSTRO is the first method t
o directly estimate 3D body-scene contact from a single image. We demonstrate th
at BSTRO significantly outperforms the prior art. The code and dataset will be a
vailable for research purposes.
****

Advancing High-Resolution Video-Language Representation With Large-Scale Video T
ranscriptions
Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlo
ng Fu, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision an

d Pattern Recognition (CVPR), 2022, pp. 5036-5045
We study joint video and language (VL) pre-training to enable cross-modality lea rning and benefit plentiful downstream VL tasks. Existing works either extract l ow-quality video features or learn limited text embedding, while neglecting that high-resolution videos and diversified semantics can significantly improve cros s-modality learning. In this paper, we propose a novel High-resolution and Diver sified VIdeo-LAnguage pre-training model (HD-VILA) for many visual tasks. In par ticular, we collect a large dataset with two distinct properties: 1) the first h igh-resolution dataset including 371.5k hours of 720p videos, and 2) the most di versified dataset covering 15 popular YouTube categories. To enable VL pre-train ing, we jointly optimize the HD-VILA model by a hybrid Transformer that learns r ich spatiotemporal features, and a multimodal Transformer that enforces interact ions of the learned video features with diversified texts. Our pre-training mode l achieves new state-of-the-art results in 10 VL understanding tasks and 2 more novel text-to-visual generation tasks. For example, we outperform SOTA models wi th relative increases of 40.4% R@1 in zero-shot MSR-VTT text-to-video retrieval task, and 55.4% in high-resolution dataset LSMDC. The learned VL embedding is al so effective in generating visually pleasing and semantically relevant results i n text-to-visual editing and super-resolution tasks.
****

Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Bench mark To Fuse Infrared and Visible for Object Detection
Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, Zhongxua n Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 5802-5811
This study addresses the issue of fusing infrared and visible images that appear differently for object detection. Aiming at generating an image of high visual quality, previous approaches discover commons underlying the two modalities and fuse upon the common space either by iterative optimization or deep networks. Th ese approaches neglect that modality differences implying the complementary info rmation are extremely important for both fusion and subsequent detection task. T his paper proposes a bilevel optimization formulation for the joint problem of f usion and detection, and then unrolls to a target-aware Dual Adversarial Learnin g (TarDAL) network for fusion and a commonly used detection network. The fusion network with one generator and dual discriminators seeks commons while learning from differences, which preserves structural information of targets from the inf rared and textural details from the visible. Furthermore, we build a synchronize d imaging system with calibrated infrared and optical sensors, and collect curre ntly the most comprehensive benchmark covering a wide range of scenarios. Extens ive experiments on several public datasets and our benchmark demonstrate that ou r method outputs not only visually appealing fusion but also higher detection mA P than the state-of-the-art approaches.
****

En-Compactness: Self-Distillation Embedding & Contrastive Generation for General ized Zero-Shot Learning
Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, Yanyun Qu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9306-9315
Generalized zero-shot learning (GZSL) requires a classifier trained on seen clas ses that can recognize objects from both seen and unseen classes. Due to the abs ence of unseen training samples, the classifier tends to bias towards seen class es. To mitigate this problem, feature generation based models are proposed to sy nthesize visual features for unseen classes. However, these features are generat ed in the visual feature space which lacks of discriminative ability. Therefore, some methods turn to find a better embedding space for the classifier training. They emphasize the inter-class relationships of seen classes, leading the embed ding space overfitted to seen classes and unfriendly to unseen classes. Instead, in this paper, we propose an Intra-Class Compactness Enhancement method (ICCE) for GZSL. Our ICCE promotes intra-class compactness with inter-class separabilit y on both seen and unseen classes in the embedding space and visual feature spac

e. By promoting the intra-class relationships but the inter-class structures, we can distinguish different classes with better generalization. Specifically, we propose a Self-Distillation Embedding (SDE) module and a Semantic-Visual Contrastive Generation (SVCG) module. The former promotes intra-class compactness in the embedding space, while the latter accomplishes it in the visual feature space. The experiments demonstrate that our ICCE outperforms the state-of-the-art methods on four datasets and achieves competitive results on the remaining dataset.
****

Neural Face Identification in a 2D Wireframe Projection of a Manifold Object
Kehan Wang, Jia Zheng, Zihan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1622-1631
In computer-aided design (CAD) systems, 2D line drawings are commonly used to illustrate 3D object designs. To reconstruct the 3D models depicted by a single 2D line drawing, an important key is finding the edge loops in the line drawing which correspond to the actual faces of the 3D object. In this paper, we approach the classical problem of face identification from a novel data-driven point of view. We cast it as a sequence generation problem: starting from an arbitrary edge, we adopt a variant of the popular Transformer model to predict the edges associated with the same face in a natural order. This allows us to avoid searching the space of all possible edge loops with various hand-crafted rules and heuristics as most existing methods do, deal with challenging cases such as curved surfaces and nested edge loops, and leverage additional cues such as face types. We further discuss how possibly imperfect predictions can be used for 3D object reconstruction. The project page is at https://manycore-research.github.io/faceformer.
****

LC-FDNet: Learned Lossless Image Compression With Frequency Decomposition Network
Hochang Rhee, Yeong Il Jang, Seyun Kim, Nam Ik Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6033-6042
Recent learning-based lossless image compression methods encode an image in the unit of subimages and achieve comparable performances to conventional non-learning algorithms. However, these methods do not consider the performance drop in the high-frequency region, giving equal consideration to the low and high-frequency areas. In this paper, we propose a new lossless image compression method that proceeds the encoding in a coarse-to-fine manner to separate and process low and high-frequency regions differently. We initially compress the low-frequency components and then use them as additional input for encoding the remaining high-frequency region. The low-frequency components act as a strong prior in this case, which leads to improved estimation in the high-frequency area. In addition, we design the frequency decomposition process to be adaptive to color channel, spatial location, and image characteristics. As a result, our method derives an image-specific optimal ratio of low/high-frequency components. Experiments show that the proposed method achieves state-of-the-art performance for benchmark high-resolution datasets.
****

Nonuniform-to-Uniform Quantization: Towards Accurate Quantization via Generalized Straight-Through Estimation
Zechun Liu, Kwang-Ting Cheng, Dong Huang, Eric P. Xing, Zhiqiang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4942-4952
The nonuniform quantization strategy for compressing neural networks usually achieves better performance than its counterpart, i.e., uniform strategy, due to its superior representational capacity. However, many nonuniform quantization methods overlook the complicated projection process in implementing the nonuniformly quantized weights/activations, which incurs non-negligible time and space overhead in hardware deployment. In this study, we propose Nonuniform-to-Uniform Quantization (N2UQ), a method that can maintain the strong representation ability of nonuniform methods while being hardware-friendly and efficient as the uniform q

uantization for model inference. We achieve this through learning the flexible i
n-equidistant input thresholds to better fit the underlying distribution while q
uantizing these real-valued inputs into equidistant output levels. To train the
quantized network with learnable input thresholds, we introduce a generalized st
raight-through estimator (G-STE) for intractable backward derivative calculation
 w.r.t. threshold parameters. Additionally, we consider entropy preserving regul
arization to further reduce information loss in weight quantization. Even under
this adverse constraint of imposing uniformly quantized weights and activations,
 our N2UQ outperforms state-of-the-art nonuniform quantization methods by 0.5 1.
7% on ImageNet, demonstrating the contribution of N2UQ design. Code and models a
re available at: https://github.com/liuzechun/Nonuniform-to-Uniform-Quantization
.
****

Deep Rectangling for Image Stitching: A Learning Baseline
Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, Yao Zhao; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5
740-5748
Stitched images provide a wide field-of-view (FoV) but suffer from unpleasant ir
regular boundaries. To deal with this problem, existing image rectangling method
s devote to searching an initial mesh and optimizing a target mesh to form the m
esh deformation in two stages. Then rectangular images can be generated by warpi
ng stitched images. However, these solutions only work for images with rich line
ar structures, leading to noticeable distortions for portraits and landscapes wi
th non-linear objects. In this paper, we address these issues by proposing the f
irst deep learning solution to image rectangling. Concretely, we predefine a rig
id target mesh and only estimate an initial mesh to form the mesh deformation, c
ontributing to a compact one-stage solution. The initial mesh is predicted using
 a fully convolutional network with a residual progressive regression strategy.
To obtain results with high content fidelity, a comprehensive objective function
 is proposed to simultaneously encourage the boundary rectangular, mesh shape-pr
eserving, and content perceptually natural. Besides, we build the first image st
itching rectangling dataset with a large diversity in irregular boundaries and s
cenes. Extensive experiments demonstrate our superiority over traditional method
s both quantitatively and qualitatively. The codes and dataset will be available
.
****

PCL: Proxy-Based Contrastive Learning for Domain Generalization
Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, B
ei Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2022, pp. 7097-7107
Domain generalization refers to the problem of training a model from a collectio
n of different source domains that can directly generalize to the unseen target
domains. A promising solution is contrastive learning, which attempts to learn d
omain-invariant representations by exploiting rich semantic relations among samp
le-to-sample pairs from different domains. A simple approach is to pull positive
 sample pairs from different domains closer while pushing other negative pairs f
urther apart. In this paper, we find that directly applying contrastive-based me
thods (e.g., supervised contrastive learning) are not effective in domain genera
lization. We argue that aligning positive sample-to-sample pairs tends to hinder
 the model generalization due to the significant distribution gaps between diffe
rent domains. To address this issue, we propose a novel proxy-based contrastive
learning method, which replaces the original sample-to-sample relations with pro
xy-to-sample relations, significantly alleviating the positive alignment issue.
Experiments on the four standard benchmarks demonstrate the effectiveness of the
 proposed method. Furthermore, we also consider a more complex scenario where no
 ImageNet pre-trained models are provided. Our method consistently shows better
performance.
****

SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estim
ation With Learnt Surface Embeddings

Rasmus Laurvig Haugaard, Anders Glent Buch; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6749-6758
We present an approach to learn dense, continuous 2D-3D correspondence distribut
ions over the surface of objects from data with no prior knowledge of visual amb
iguities like symmetry. We also present a new method for 6D pose estimation of r
igid objects using the learnt distributions to sample, score and refine pose hyp
otheses. The correspondence distributions are learnt with a contrastive loss, re
presented in object-specific latent spaces by an encoder-decoder query model and
 a small fully connected key model. Our method is unsupervised with respect to v
isual ambiguities, yet we show that the query- and key models learn to represent
 accurate multi-modal surface distributions. Our pose estimation method improves
 the state-of-the-art significantly on the comprehensive BOP Challenge, trained
purely on synthetic data, even compared with methods trained on real data. The p
roject site is at surfemb.github.io.
****

Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Cre
ation
Naofumi Akimoto, Yuhi Matsuo, Yoshimitsu Aoki; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11441-11450
We address the problem of generating a 360-degree image from a single image with
 a narrow field of view by estimating its surroundings. Previous methods suffere
d from overfitting to the training resolution and deterministic generation. This
 paper proposes a completion method using a transformer for scene modeling and n
ovel methods to improve the properties of a 360-degree image on the output image
. Specifically, we use CompletionNets with a transformer to perform diverse comp
letions and AdjustmentNet to match color, stitching, and resolution with an inpu
t image, enabling inference at any resolution. To improve the properties of a 36
0-degree image on an output image, we also propose WS-perceptual loss and circul
ar inference. Thorough experiments show that our method outperforms state-of-the
-art (SOTA) methods both qualitatively and quantitatively. For example, compared
 to SOTA methods, our method completes images 16 times larger in resolution and
achieves 1.7 times lower Frechet inception distance (FID). Furthermore, we propo
se a pipeline that uses the completion results for lighting and background of 3D
CG scenes. Our plausible background completion enables perceptually natural resu
lts in the application of inserting virtual objects with specular surfaces.
****

Learning 3D Object Shape and Layout Without 3D Supervision
Georgia Gkioxari, Nikhila Ravi, Justin Johnson; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1695-1704
A 3D scene consists of a set of objects, each with a shape and a layout giving t
heir position in space. Understanding 3D scenes from 2D images is an important g
oal, with applications in robotics and graphics. While there have been recent ad
vances in predicting 3D shape and layout from a single image, most approaches re
ly on 3D ground truth for training which is expensive to collect at scale. We ov
ercome these limitations and propose a method that learns to predict 3D shape an
d layout for objects without any ground truth shape or layout information: inste
ad we rely on multi-view images with 2D supervision which can more easily be col
lected at scale. Through extensive experiments on ShapeNet, Hypersim, and ScanNe
t we demonstrate that our approach scales to large datasets of realistic images,
 and compares favorably to methods relying on 3D ground truth. On Hypersim and S
canNet where reliable 3D ground truth is not available, our approach outperforms
 supervised approaches trained on smaller and less diverse datasets.
****

An Empirical Study of End-to-End Temporal Action Detection
Xiaolong Liu, Song Bai, Xiang Bai; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2022, pp. 20010-20019
Temporal action detection (TAD) is an important yet challenging task in video un
derstanding. It aims to simultaneously predict the semantic label and the tempor
al interval of every action instance in an untrimmed video. Rather than end-to-e
nd learning, most existing methods adopt a head-only learning paradigm, where th

e video encoder is pre-trained for action classification, and only the detection head upon the encoder is optimized for TAD. The effect of end-to-end learning is not systematically evaluated. Besides, there lacks an in-depth study on the efficiency-accuracy trade-off in end-to-end TAD. In this paper, we present an empirical study of end-to-end temporal action detection. We validate the advantage of end-to-end learning over head-only learning and observe up to 11% performance improvement. Besides, we study the effects of multiple design choices that affect the TAD performance and speed, including detection head, video encoder, and resolution of input videos. Based on the findings, we build a mid-resolution baseline detector, which achieves the state-of-the-art performance of end-to-end methods while running more than 4x faster. We hope that this paper can serve as a guide for end-to-end learning and inspire future research in this field. Code and models are available at https://github.com/xlliu7/E2E-TAD.

****

SimVP: Simpler Yet Better Video Prediction

Zhangyang Gao, Cheng Tan, Lirong Wu, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3170-3180

From CNN, RNN, to ViT, we have witnessed remarkable advancements in video prediction, incorporating auxiliary inputs, elaborate neural architectures, and sophisticated training strategies. We admire these progresses but are confused about the necessity: is there a simple method that can perform comparably well? This paper proposes SimVP, a simple video prediction model that is completely built upon CNN and trained by MSE loss in an end-to-end fashion. Without introducing any additional tricks and complicated strategies, we can achieve state-of-the-art performance on five benchmark datasets. Through extended experiments, we demonstrate that SimVP has strong generalization and extensibility on real-world datasets. The significant reduction of training cost makes it easier to scale to complex scenarios. We believe SimVP can serve as a solid baseline to stimulate the further development of video prediction.

****

Object Localization Under Single Coarse Point Supervision

Xuehui Yu, Pengfei Chen, Di Wu, Najmul Hassan, Guorong Li, Junchi Yan, Humphrey Shi, Qixiang Ye, Zhenjun Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4868-4877

Point-based object localization (POL), which pursues high-performance object sensing under low-cost data annotation, has attracted increased attention. However, the point annotation mode inevitably introduces semantic variance for the inconsistency of annotated points. Existing POL methods heavily reply on accurate key-point annotations which are difficult to define. In this study, we propose a POL method using coarse point annotations, relaxing the supervision signals from accurate key points to freely spotted points. To this end, we propose a coarse point refinement (CPR) approach, which to our best knowledge is the first attempt to alleviate semantic variance from the perspective of algorithm. CPR constructs point bags, selects semantic-correlated points, and produces semantic center points through multiple instance learning (MIL). In this way, CPR defines a weakly supervised evolution procedure, which ensures training high-performance object localizer under coarse point supervision. Experimental results on COCO, DOTA and our proposed SeaPerson dataset validate the effectiveness of the CPR approach. The dataset and code will be available at https://github.com/ucas-vg/PointTinyBenchmark/

****

Unsupervised Learning of Accurate Siamese Tracking

Qiuhong Shen, Lei Qiao, Jinyang Guo, Peixia Li, Xin Li, Bo Li, Weitao Feng, Weihao Gan, Wei Wu, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8101-8110

Unsupervised learning has been popular in various computer vision tasks, including visual object tracking. However, prior unsupervised tracking approaches rely heavily on spatial supervision from template-search pairs and are still unable to track objects with strong variation over a long time span. As unlimited self-supervision signals can be obtained by tracking a video along a cycle in time, we

investigate evolving a Siamese tracker by tracking videos forward-backward. We present a novel unsupervised tracking framework, in which we can learn temporal correspondence both on the classification branch and regression branch. Specifically, to propagate reliable template feature in the forward propagation process so that the tracker can be trained in the cycle, we first propose a consistency propagation transformation. We then identify an ill-posed penalty problem in conventional cycle training in backward propagation process. Thus, a differentiable region mask is proposed to select features as well as to implicitly penalize tracking errors on intermediate frames. Moreover, since noisy labels may degrade training, we propose a mask-guided loss reweighting strategy to assign dynamic weights based on the quality of pseudo labels. In extensive experiments, our tracker outperforms preceding unsupervised methods by a substantial margin, performing on par with supervised methods on large-scale datasets such as TrackingNet and LaSOT. Code is available at https://github.com/FlorinShum/ULAST.

****

Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection

Hitesh Sapkota, Qi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3212-3221

Multiple-instance learning (MIL) provides an effective way to tackle the video anomaly detection problem by modeling it as a weakly supervised problem as the labels are usually only available at the video level while missing for frames due to expensive labeling cost. We propose to conduct novel Bayesian non-parametric submodular video partition (BN-SVP) to significantly improve MIL model training that can offer a highly reliable solution for robust anomaly detection in practical settings that include outlier segments or multiple types of abnormal events. BN-SVP essentially performs dynamic non-parametric hierarchical clustering with an enhanced self-transition that groups segments in a video into temporally consistent and semantically coherent hidden states that can be naturally interpreted as scenes. Each segment is assumed to be generated through a non-parametric mixture process that allows variations of segments within the same scenes to accommodate the dynamic and noisy nature of many real-world surveillance videos. The scene and mixture component assignment of BN-SVP also induces a pairwise similarity among segments, resulting in non-parametric construction of a submodular set function. Integrating this function with an MIL loss effectively exposes the model to a diverse set of potentially positive instances to improve its training. A greedy algorithm is developed to optimize the submodular function and support efficient model training. Our theoretical analysis ensures a strong performance guarantee of the proposed algorithm. The effectiveness of the proposed approach is demonstrated over multiple real-world anomaly video datasets with robust detection performance.

****

Brain-Supervised Image Editing

Keith M. Davis III, Carlos de la Torre-Ortiz, Tuukka Ruotsalo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18480-18489

Despite recent advances in deep neural models for semantic image editing, present approaches are dependent on explicit human input. Previous work assumes the availability of manually curated datasets for supervised learning, while for unsupervised approaches the human inspection of discovered components is required to identify those which modify worthwhile semantic features. Here, we present a novel alternative: the utilization of brain responses as a supervision signal for learning semantic feature representations. Participants (N=30) in a neurophysiological experiment were shown artificially generated faces and instructed to look for a particular semantic feature, such as "old" or "smiling", while their brain responses were recorded via electroencephalography (EEG). Using supervision signals inferred from these responses, semantic features within the latent space of a generative adversarial network (GAN) were learned and then used to edit semantic features of new images. We show that implicit brain supervision achieves comparable semantic image editing performance to explicit manual labeling. This work demonstrates the feasibility of utilizing implicit human reactions recorded vi

a brain-computer interfaces for semantic image editing and interpretation.
****
3D Shape Variational Autoencoder Latent Disentanglement via Mini-Batch Feature S
wapping for Bodies and Faces
Simone Foti, Bongjin Koo, Danail Stoyanov, Matthew J. Clarkson; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
pp. 18730-18739
Learning a disentangled, interpretable, and structured latent representation in
3D generative models of faces and bodies is still an open problem. The problem i
s particularly acute when control over identity features is required. In this pa
per, we propose an intuitive yet effective self-supervised approach to train a 3
D shape variational autoencoder (VAE) which encourages a disentangled latent rep
resentation of identity features. Curating the mini-batch generation by swapping
 arbitrary features across different shapes allows to define a loss function lev
eraging known differences and similarities in the latent representations. Experi
mental results conducted on 3D meshes show that state-of-the-art methods for lat
ent disentanglement are not able to disentangle identity features of faces and b
odies. Our proposed method properly decouples the generation of such features wh
ile maintaining good representation and reconstruction capabilities. Our code an
d pre-trained models are available at github.com/simofoti/3DVAE-SwapDisentangled
.
****
Unified Transformer Tracker for Object Tracking
Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, Zhicheng Yan
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 8781-8790
As an important area in computer vision, object tracking has formed two separate
 communities that respectively study Single Object Tracking (SOT) and Multiple O
bject Tracking (MOT). However, current methods in one tracking scenario are not
easily adapted to the other due to the divergent training datasets and tracking
objects of both tasks. Although UniTrack demonstrates that a shared appearance m
odel with multiple heads can be used to tackle individual tracking tasks, it fai
ls to exploit the large-scale tracking datasets for training and performs poorly
 on single object tracking. In this work, we present the Unified Transformer Tra
cker (UTT) to address tracking problems in different scenarios with one paradigm
. A track transformer is developed in our UTT to track the target in both SOT an
d MOT where the correlation between the target feature and the tracking frame fe
ature is exploited to localize the target. We demonstrate that both SOT and MOT
tasks can be solved within this framework, and the model can be simultaneously e
nd-to-end trained by alternatively optimizing the SOT and MOT objectives on the
datasets of individual tasks. Extensive experiments are conducted on several ben
chmarks with a unified model trained on both SOT and MOT datasets.
****
Non-Parametric Depth Distribution Modelling Based Depth Inference for Multi-View
 Stereo
Jiayu Yang, Jose M. Alvarez, Miaomiao Liu; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8626-8634
Recent cost volume pyramid based deep neural networks have unlocked the potentia
l of efficiently leveraging high-resolution images for depth inference from mult
i-view stereo. In general, those approaches assume that the depth of each pixel
follows a unimodal distribution. Boundary pixels usually follow a multi-modal di
stribution as they represent different depths; Therefore, the assumption results
 in an erroneous depth prediction at the coarser level of the cost volume pyrami
d and can not be corrected in the refinement levels leading to wrong depth predi
ctions. In contrast, we propose constructing the cost volume by non-parametric d
epth distribution modeling to handle pixels with unimodal and multi-modal distri
butions. Our approach outputs multiple depth hypotheses at the coarser level to
avoid errors in the early stage. As we perform local search around these multipl
e hypotheses in subsequent levels, our approach does not maintain the rigid dept
h spatial ordering and, therefore, we introduce a sparse cost aggregation networ

k to derive information within each volume. We evaluate our approach extensively on two benchmark datasets: DTU and Tanks & Temples. Our experimental results show that our model outperforms existing methods by a large margin and achieves superior performance on boundary regions. Code is available at https://github.com/NVlabs/NP-CVP-MVSNet

****

Equalized Focal Loss for Dense Long-Tailed Object Detection

Bo Li, Yongqiang Yao, Jingru Tan, Gang Zhang, Fengwei Yu, Jianwei Lu, Ye Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6990-6999

Despite the recent success of long-tailed object detection, almost all long-tailed object detectors are developed based on the two-stage paradigm. In practice, one-stage detectors are more prevalent in the industry because they have a simple and fast pipeline that is easy to deploy. However, in the long-tailed scenario, this line of work has not been explored so far. In this paper, we investigate whether one-stage detectors can perform well in this case. We discover the primary obstacle that prevents one-stage detectors from achieving excellent performance is: categories suffer from different degrees of positive-negative imbalance problems under the long-tailed data distribution. The conventional focal loss balances the training process with the same modulating factor for all categories, thus failing to handle the long-tailed problem. To address this issue, we propose the Equalized Focal Loss (EFL) that rebalances the loss contribution of positive and negative samples of different categories independently according to their imbalance degrees. Specifically, EFL adopts a category-relevant modulating factor which can be adjusted dynamically by the training status of different categories. Extensive experiments conducted on the challenging LVIS v1 benchmark demonstrate the effectiveness of our proposed method. With an end-to-end training pipeline, EFL achieves 29.2% in terms of overall AP and obtains significant performance improvements on rare categories, surpassing all existing state-of-the-art methods. The code is available at https://github.com/ModelTC/EOD.

****

Generating High Fidelity Data From Low-Density Regions Using Diffusion Models

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, Cristian Canton; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11492-11501

Our work focuses on addressing sample deficiency from low-density regions of data manifold in common image datasets. We leverage diffusion process based generative models to synthesize novel images from low-density regions. We observe that uniform sampling from diffusion models predominantly samples from high-density regions of the data manifold. Therefore, we modify the sampling process to guide it towards low-density regions while simultaneously maintaining the fidelity of synthetic data. We rigorously demonstrate that our process successfully generates novel high fidelity samples from low-density regions. We further examine generated samples and show that the model does not memorize low-density data and indeed learns to generate novel samples from low-density regions.

****

DeepDPM: Deep Clustering With an Unknown Number of Clusters

Meitar Ronen, Shahaf E. Finder, Oren Freifeld; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9861-9870

Deep Learning (DL) has shown great promise in the unsupervised task of clustering. That said, while in classical (i.e., non-deep) clustering the benefits of the nonparametric approach are well known, most deep-clustering methods are parametric: namely, they require a predefined and fixed number of clusters, denoted by K. When K is unknown, however, using model-selection criteria to choose its optimal value might become computationally expensive, especially in DL as the training process would have to be repeated numerous times. In this work, we bridge this gap by introducing an effective deep-clustering method that does not require knowing the value of K as it infers it during the learning. Using a split/merge framework, a dynamic architecture that adapts to the changing K, and a novel loss, our proposed method outperforms existing nonparametric methods (both classical

and deep ones). While the very few existing deep nonparametric methods lack sca
lability, we demonstrate ours by being the first to report the performance of su
ch a method on ImageNet. We also demonstrate the importance of inferring K by sh
owing how methods that fix it deteriorate in performance when their assumed K va
lue gets further from the ground-truth one, especially on imbalanced datasets. O
ur code is available at https://github.com/BGU-CS-VIL/DeepDPM.
****

Spiking Transformers for Event-Based Single Object Tracking
Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, Xi
n Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2022, pp. 8801-8810
Event-based cameras bring a unique capability to tracking, being able to functio
n in challenging real-world conditions as a direct result of their high temporal
 resolution and high dynamic range. These imagers capture events asynchronously
that encode rich temporal and spatial information. However, effectively extracti
ng this information from events remains an open challenge. In this work, we prop
ose a spiking transformer network, STNet, for single object tracking. STNet dyna
mically extracts and fuses information from both temporal and spatial domains. I
n particular, the proposed architecture features a transformer module to provide
 global spatial information and a spiking neural network (SNN) module for extrac
ting temporal cues. The spiking threshold of the SNN module is dynamically adjus
ted based on the statistical cues of the spatial information, which we find esse
ntial in providing robust SNN features. We fuse both feature branches dynamicall
y with a novel cross-domain attention fusion algorithm. Extensive experiments on
 three event-based datasets, FE240hz, EED and VisEvent validate that the propose
d STNet outperforms existing state-of-the-art methods in both tracking accuracy
and speed with a significant margin.
****

FocalClick: Towards Practical Interactive Image Segmentation
Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, Hengshuang Zhao; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2022, pp. 1300-1309
Interactive segmentation allows users to extract target masks by making positive
/negative clicks. Although explored by many previous works, there is still a gap
 between academic approaches and industrial needs: first, existing models are no
t efficient enough to work on low power devices; second, they perform poorly whe
n used to refine preexisting masks as they could not avoid destroying the correc
t part. FocalClick solves both issues at once by predicting and updating the mas
k in localized areas. For higher efficiency, we decompose the slow prediction on
 the entire image into two fast inferences on small crops: a coarse segmentation
 on the Target Crop, and a local refinement on the Focus Crop. To make the model
 work with preexisting masks, we formulate a sub-task termed Interactive Mask Co
rrection, and propose Progressive Merge as the solution. Progressive Merge explo
its morphological information to decide where to preserve and where to update, e
nabling users to refine any preexisting mask effectively. FocalClick achieves co
mpetitive results against SOTA methods with significantly smaller FLOPs. It also
 shows significant superiority when making corrections on preexisting masks. Cod
e and data will be released at github.com/XavierCHEN34/ClickSEG
****

ISDNet: Integrating Shallow and Deep Networks for Efficient Ultra-High Resolutio
n Segmentation
Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Gua
nnan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, Ke Xu; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4361-4370
The huge burden of computation and memory are two obstacles in ultra-high resolu
tion image segmentation. To tackle these issues, most of the previous works foll
ow the global-local refinement pipeline, which pays more attention to the memory
 consumption but neglects the inference speed. In comparison to the pipeline tha
t partitions the large image into small local regions, we focus on inferring the
 whole image directly. In this paper, we propose ISDNet, a novel ultra-high reso

lution segmentation framework that integrates the shallow and deep networks in a new manner, which significantly accelerates the inference speed while achieving accurate segmentation. To further exploit the relationship between the shallow and deep features, we propose a novel Relational-Aware feature Fusion module, which ensures high performance and robustness of our framework. Extensive experiments on Deepglobe, Inria Aerial, and Cityscapes datasets demonstrate our performance is consistently superior to state-of-the-arts. Specifically, it achieves 73.30 mIoU with a speed of 27.70 FPS on Deepglobe, which is more accurate and 172 x faster than the recent competitor. Code available at https://github.com/cedricgsh/ISDNet.

****

Unsupervised Domain Adaptation for Nighttime Aerial Tracking

Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, Guang Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8896-8905

Previous advances in object tracking mostly reported on favorable illumination circumstances while neglecting performance at nighttime, which significantly impeded the development of related aerial robot applications. This work instead develops a novel unsupervised domain adaptation framework for nighttime aerial tracking (named UDAT). Specifically, a unique object discovery approach is provided to generate training patches from raw nighttime tracking videos. To tackle the domain discrepancy, we employ a Transformer-based bridging layer post to the feature extractor to align image features from both domains. With a Transformer day/night feature discriminator, the daytime tracking model is adversarially trained to track at night. Moreover, we construct a pioneering benchmark namely NAT2021 for unsupervised domain adaptive nighttime tracking, which comprises a test set of 180 manually annotated tracking sequences and a train set of over 276k unlabelled nighttime tracking frames. Exhaustive experiments demonstrate the robustness and domain adaptability of the proposed framework in nighttime aerial tracking. The code and benchmark are available at https://github.com/vision4robotics/UDAT.

****

Balanced Multimodal Learning via On-the-Fly Gradient Modulation

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, Di Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8238-8247

Audio-visual learning helps to comprehensively understand the world, by integrating different senses. Accordingly, multiple input modalities are expected to boost model performance, but we actually find that they are not fully exploited even when the multi-modal model outperforms its uni-modal counterpart. Specifically, in this paper we point out that existing audio-visual discriminative models, in which uniform objective is designed for all modalities, could remain under-optimized uni-modal representations, caused by another dominated modality in some scenarios, e.g., sound in blowing wind event, vision in drawing picture event, etc. To alleviate this optimization imbalance, we propose on-the-fly gradient modulation to adaptively control the optimization of each modality, via monitoring the discrepancy of their contribution towards the learning objective. Further, an extra Gaussian noise that changes dynamically is introduced to avoid possible generalization drop caused by gradient modulation. As a result, we achieve considerable improvement over common fusion methods on different audio-visual tasks, and this simple strategy can also boost existing multi-modal methods, which illustrates its efficacy and versatility.

****

RestoreFormer: High-Quality Blind Face Restoration From Undegraded Key-Value Pairs

Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17512-17521

Blind face restoration is to recover a high-quality face image from unknown degradations. As face image contains abundant contextual information, we propose a m

ethod, RestoreFormer, which explores fully-spatial attentions to model contextua
l information and surpasses existing works that use local convolutions. RestoreF
ormer has several benefits compared to prior arts. First, unlike the conventiona
l multi-head self-attention in previous Vision Transformers (ViTs), RestoreForme
r incorporates a multi-head cross-attention layer to learn fully-spatial interac
tions between corrupted queries and high-quality key-value pairs. Second, the ke
y-value pairs in ResotreFormer are sampled from a reconstruction-oriented high-q
uality dictionary, whose elements are rich in high-quality facial features speci
fically aimed for face reconstruction, leading to superior restoration results.
Third, RestoreFormer outperforms advanced state-of-the-art methods on one synthe
tic dataset and three real-world datasets, as well as produces images with bette
r visual quality. Code is available at https://github.com/wzhouxiff/RestoreForme
r.git.
****

Understanding Uncertainty Maps in Vision With Statistical Testing
Jurijs Nazarovs, Zhichun Huang, Songwong Tasneeyapant, Rudrasis Chakraborty, Vik
as Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2022, pp. 406-416
Quantitative descriptions of confidence intervals and uncertainties of the predi
ctions of a model are needed in many applications in vision and machine learning
. Mechanisms that enable this for deep neural network (DNN) models are slowly be
coming available, and occasionally, being integrated within production systems.
But the literature is sparse in terms of how to perform statistical tests with t
he uncertainties produced by these overparameterized models. For two models with
 a similar accuracy profile, is the former model's uncertainty behavior better i
n a statistically significant sense compared to the second model? For high resol
ution images, performing hypothesis tests to generate meaningful actionable info
rmation (say, at a user specified significance level 0.05) is difficult but need
ed in both mission critical settings and elsewhere. In this paper, specifically
for uncertainties defined on images, we show how revisiting results from Random
Field theory (RFT) when paired with DNN tools (to get around computational hurdl
es) leads to efficient frameworks that can provide a hypothesis test capabilitie
s, not otherwise available, for uncertainty maps from models used in many vision
 tasks. We show via many different experiments the viability of this framework.
****

CAFE: Learning To Condense Dataset by Aligning Features
Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Ha
kan Bilen, Xinchao Wang, Yang You; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2022, pp. 12196-12205
Dataset condensation aims at reducing the network training effort through conden
sing a cumbersome training set into a compact synthetic one. State-of-the-art ap
proaches largely rely on learning the synthetic data by matching the gradients b
etween the real and synthetic data batches. Despite the intuitive motivation and
 promising results, such gradient-based methods, by nature, easily overfit to a
biased set of samples that produce dominant gradients, and thus lack a global su
pervision of data distribution. In this paper, we propose a novel scheme to Cond
ense dataset by Aligning FEatures (CAFE), which explicitly attempts to preserve
the real-feature distribution as well as the discriminant power of the resulting
 synthetic set, lending itself to strong generalization capability to various ar
chitectures. At the heart of our approach is an effective strategy to align feat
ures from the real and synthetic data across various scales, while accounting fo
r the classification of real samples. Our scheme is further backed up by a novel
 dynamic bi-level optimization, which adaptively adjusts parameter updates to pr
event over-/under-fitting. We validate the proposed CAFE across various datasets
, and demonstrate that it generally outperforms the state of the art: on the SVH
N dataset, for example, the performance gain is up to 11%. Extensive experiments
 and analysis verify the effectiveness and necessity of proposed designs.
****

Causality Inspired Representation Learning for Domain Generalization
Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, Di Liu

; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 8046-8056
Domain generalization (DG) is essentially an out-of-distribution problem, aiming
 to generalize the knowledge learned from multiple source domains to an unseen t
arget domain. The mainstream is to leverage statistical models to model the depe
ndence between data and labels, intending to learn representations independent o
f domain. Nevertheless, the statistical models are superficial descriptions of r
eality since they are only required to model dependence instead of the intrinsic
 causal mechanism. When the dependence changes with the target distribution, the
 statistic models may fail to generalize. In this regard, we introduce a general
 structural causal model to formalize the DG problem. Specifically, we assume th
at each input is constructed from a mix of causal factors (whose relationship wi
th the label is invariant across domains) and non-causal factors (category-indep
endent), and only the former cause the classification judgments. Our goal is to
extract the causal factors from inputs and then reconstruct the invariant causal
 mechanisms. However, the theoretical idea is far from practical of DG since the
 required causal/non-causal factors are unobserved. We highlight that ideal caus
al factors should meet three basic properties: separated from the non-causal one
s, jointly independent, and causally sufficient for the classification. Based on
 that, we propose a Causality Inspired Representation Learning (CIRL) algorithm
that enforces the representation to satisfy the above properties and then uses t
hem to simulate the causal factors, which yields improved generalization ability
. Extensive experimental results on several widely used datasets verify the effe
ctiveness of our approach.
****

Mask-Guided Spectral-Wise Transformer for Efficient Hyperspectral Image Reconstr
uction
Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Tim
ofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2022, pp. 17502-17511
Hyperspectral image (HSI) reconstruction aims to recover the 3D spatial-spectral
 signal from a 2D measurement in the coded aperture snapshot spectral imaging (C
ASSI) system. The HSI representations are highly similar and correlated across t
he spectral dimension. Modeling the inter-spectra interactions is beneficial for
 HSI reconstruction. However, existing CNN-based methods show limitations in cap
turing spectral-wise similarity and long-range dependencies. Besides, the HSI in
formation is modulated by a coded aperture (physical mask) in CASSI. Nonetheless
, current algorithms have not fully explored the guidance effect of the mask for
 HSI restoration. In this paper, we propose a novel framework, Mask-guided Spect
ral-wise Transformer (MST), for HSI reconstruction. Specifically, we present a S
pectral-wise Multi-head Self-Attention (S-MSA) that treats each spectral feature
 as a token and calculates self-attention along the spectral dimension. In addit
ion, we customize a Mask-guided Mechanism (MM) that directs S-MSA to pay attenti
on to spatial regions with high-fidelity spectral representations. Extensive exp
eriments show that our MST significantly outperforms state-of-the-art (SOTA) met
hods on simulation and real HSI datasets while requiring dramatically cheaper co
mputational and memory costs. https://github.com/caiyuanhao1998/MST/
****

A Variational Bayesian Method for Similarity Learning in Non-Rigid Image Registr
ation
Daniel Grzech, Mohammad Farid Azampour, Ben Glocker, Julia Schnabel, Nassir Nava
b, Bernhard Kainz, Loïc Le Folgoc; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2022, pp. 119-128
We propose a novel variational Bayesian formulation for diffeomorphic non-rigid
registration of medical images, which learns in an unsupervised way a data-speci
fic similarity metric. The proposed framework is general and may be used togethe
r with many existing image registration models. We evaluate it on brain MRI scan
s from the UK Biobank and show that use of the learnt similarity metric, which i
s parametrised as a neural network, leads to more accurate results than use of t
raditional functions, e.g. SSD and LCC, to which we initialise the model, withou

t a negative impact on image registration speed or transformation smoothness. In addition, the method estimates the uncertainty associated with the transformation. The code and the trained models are available in a public repository: https://github.com/dgrzech/learnsim.

****

Not Just Selection, but Exploration: Online Class-Incremental Continual Learning via Dual View Consistency

Yanan Gu, Xu Yang, Kun Wei, Cheng Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7442-7451

Online class-incremental continual learning aims to learn new classes continually from a never-ending and single-pass data stream, while not forgetting the learned knowledge of old classes. Existing replay-based methods have shown promising performance by storing a subset of old class data. Unfortunately, these methods only focus on selecting samples from the memory bank for replay and ignore the adequate exploration of semantic information in the single-pass data stream, leading to poor classification accuracy. In this paper, we propose a novel yet effective framework for online class-incremental continual learning, which considers not only the selection of stored samples, but also the full exploration of the data stream. Specifically, we propose a gradient-based sample selection strategy, which selects the stored samples whose gradients generated in the network are most interfered by the new incoming samples. We believe such samples are beneficial for updating the neural network based on back gradient propagation. More importantly, we seek to explore the semantic information between two different views of training images by maximizing their mutual information, which is conducive to the improvement of classification accuracy. Extensive experimental results demonstrate that our method achieves state-of-the-art performance on a variety of benchmark datasets. Our code is available on https://github.com/YananGu/DVC.

****

PPDL: Predicate Probability Distribution Based Loss for Unbiased Scene Graph Generation

Wei Li, Haiwei Zhang, Qijie Bai, Guoqing Zhao, Ning Jiang, Xiaojie Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19447-19456

Scene Graph Generation (SGG) has attracted more and more attention from visual researchers in recent years, since Scene Graph (SG) is valuable in many downstream tasks due to its rich structural-semantic details. However, the application value of SG on downstream tasks is severely limited by the predicate classification bias, which is caused by long-tailed data and presented as semantic bias of predicted relation predicates. Existing methods mainly reduce the prediction bias by better aggregating contexts and integrating external priori knowledge, but rarely take the semantic similarities between predicates into account. In this paper, we propose a Predicate Probability Distribution based Loss (PPDL) to train the biased SGG models and obtain unbiased Scene Graphs ultimately. Firstly, we propose a predicate probability distribution as the semantic representation of a particular predicate class. Afterwards, we re-balance the biased training loss according to the similarity between the predicted probability distribution and the estimated one, and eventually eliminate the long-tailed bias on predicate classification. Notably, the PPDL training method is model-agnostic, and extensive experiments and qualitative analyses on the Visual Genome dataset reveal significant performance improvements of our method on tail classes compared to the state-of-the-art methods.

****

Block-NeRF: Scalable Large Scene Neural View Synthesis

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, Henrik Kretzschmar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8248-8258

We present Block-NeRF, a variant of Neural Radiance Fields that can represent large-scale environments. Specifically, we demonstrate that when scaling NeRF to render city-scale scenes spanning multiple blocks, it is vital to decompose the s

cene into individually trained NeRFs. This decomposition decouples rendering tim
e from scene size, enables rendering to scale to arbitrarily large environments,
 and allows per-block updates of the environment. We adopt several architectural
 changes to make NeRF robust to data captured over months under different enviro
nmental conditions. We add appearance embeddings, learned pose refinement, and c
ontrollable exposure to each individual NeRF, and introduce a procedure for alig
ning appearance between adjacent NeRFs so that they can be seamlessly combined.
We build a grid of Block-NeRFs from 2.8 million images to create the largest neu
ral scene representation to date, capable of rendering an entire neighborhood of
 San Francisco.
****

Coupling Vision and Proprioception for Navigation of Legged Robots
Zipeng Fu, Ashish Kumar, Ananye Agarwal, Haozhi Qi, Jitendra Malik, Deepak Patha
k; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2022, pp. 17273-17283
We exploit the complementary strengths of vision and proprioception to develop a
 point-goal navigation system for legged robots, called VP-Nav. Legged systems a
re capable of traversing more complex terrain than wheeled robots, but to fully
utilize this capability, we need a high-level path planner in the navigation sys
tem to be aware of the walking capabilities of the low-level locomotion policy i
n varying environments. We achieve this by using proprioceptive feedback to ensu
re the safety of the planned path by sensing unexpected obstacles like glass wal
ls, terrain properties like slipperiness or softness of the ground and robot pro
perties like extra payload that are likely missed by vision. The navigation syst
em uses onboard cameras to generate an occupancy map and a corresponding cost ma
p to reach the goal. A fast marching planner then generates a target path. A vel
ocity command generator takes this as input to generate the desired velocity for
 the walking policy. A safety advisor module adds sensed unexpected obstacles to
 the occupancy map and environment-determined speed limits to the velocity comma
nd generator. We show superior performance compared to wheeled robot baselines,
and ablation studies which have disjoint high-level planning and low-level contr
ol. We also show the real-world deployment of VP-Nav on a quadruped robot with o
nboard sensors and computation. Videos at https://navigation-locomotion.github.i
o
****

Fine-Grained Predicates Learning for Scene Graph Generation
Xinyu Lyu, Lianli Gao, Yuyu Guo, Zhou Zhao, Hao Huang, Heng Tao Shen, Jingkuan S
ong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2022, pp. 19467-19475
The performance of current Scene Graph Generation models is severely hampered by
 some hard-to-distinguish predicates, e.g., "woman-on/standing on/walking on-bea
ch" or "woman-near/looking at/in front of-child". While general SGG models are p
rone to predict head predicates and existing re-balancing strategies prefer tail
 categories, none of them can appropriately handle these hard-to-distinguish pre
dicates. To tackle this issue, inspired by fine-grained image classification, wh
ich focuses on differentiating among hard-to-distinguish object classes, we prop
ose a method named Fine-Grained Predicates Learning (FGPL) which aims at differe
ntiating among hard-to-distinguish predicates for Scene Graph Generation task. S
pecifically, we first introduce a Predicate Lattice that helps SGG models to fig
ure out fine-grained predicate pairs. Then, utilizing the Predicate Lattice, we
propose a Category Discriminating Loss and an Entity Discriminating Loss, which
both contribute to distinguishing fine-grained predicates while maintaining lear
ned discriminatory power over recognizable ones. The proposed model-agnostic str
ategy significantly boosts the performances of three benchmark models (Transform
er, VCTree, and Motif) by 22.8%, 24.1% and 21.7% of Mean Recall (mR@100) on the
Predicate Classification sub-task, respectively. Our model also outperforms stat
e-of-the-art methods by a large margin (i.e., 6.1%, 4.6%, and 3.2% of Mean Recal
l (mR@100)) on the Visual Genome dataset.
****

Generalized Few-Shot Semantic Segmentation

Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, Jiaya J ia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 11563-11572

Training semantic segmentation models requires a large amount of finely annotate d data, making it hard to quickly adapt to novel classes not satisfying this con dition. Few-Shot Segmentation (FS-Seg) tackles this problem with many constraint s. In this paper, we introduce a new benchmark, called Generalized Few-Shot Sema ntic Segmentation (GFS-Seg), to analyze the generalization ability of simultaneo usly segmenting the novel categories with very few examples and the base categor ies with sufficient examples. It is the first study showing that previous repres entative state-of-the-art FS-Seg methods fall short in GFS-Seg and the performan ce discrepancy mainly comes from the constrained setting of FS-Seg. To make GFS- Seg tractable, we set up a GFS-Seg baseline that achieves decent performance wit hout structural change on the original model. Then, since context is essential f or semantic segmentation, we propose the Context-Aware Prototype Learning (CAPL) that significantly improves performance by 1) leveraging the co-occurrence prio r knowledge from support samples, and 2) dynamically enriching contextual inform ation to the classifier, conditioned on the content of each query image. Both tw o contributions are experimentally shown to have substantial practical merit. Ex tensive experiments on Pascal-VOC and COCO manifest the effectiveness of CAPL, a nd CAPL generalizes well to FS-Seg by achieving competitive performance. Code wi ll be made publicly available.

****

Exploiting Rigidity Constraints for LiDAR Scene Flow Estimation

Guanting Dong, Yueyi Zhang, Hanlin Li, Xiaoyan Sun, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 12776-12785

Previous LiDAR scene flow estimation methods, especially recurrent neural networ ks, usually suffer from structure distortion in challenging cases, such as spars e reflection and motion occlusions. In this paper, we propose a novel optimizati on method based on a recurrent neural network to predict LiDAR scene flow in a w eakly supervised manner. Specifically, our neural recurrent network exploits dir ect rigidity constraints to preserve the geometric structure of the warped sourc e scene during an iterative alignment procedure. An error awarded optimization s trategy is proposed to update the LiDAR scene flow by minimizing the point measu rement error instead of reconstructing the cost volume multiple times. Trained o n two autonomous driving datasets, our network outperforms recent state-of-the-a rt networks on lidarKITTI by a large margin. The code and models will be availab le at https://github. com/gtdong-ustc/LiDARSceneFlow.

****

Neural Head Avatars From Monocular RGB Videos

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, Justus Thies; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 18653-18664

We present Neural Head Avatars, a novel neural representation that explicitly mo dels the surface geometry and appearance of an animatable human avatar that can be used for teleconferencing in AR/VR or other applications in the movie or game s industry that rely on a digital human. Our representation can be learned from a monocular RGB portrait video that features a range of different expressions an d views. Specifically, we propose a hybrid representation consisting of a morpha ble model for the coarse shape and expressions of the face, and two feed-forward networks, predicting vertex offsets of the underlying mesh as well as a view- a nd expression-dependent texture. We demonstrate that this representation is able to accurately extrapolate to unseen poses and view points, and generates natura l expressions while providing sharp texture details. Compared to previous works on head avatars, our method provides a disentangled shape and appearance model o f the complete human head (including hair) that is compatible with the standard graphics pipeline. Moreover, it quantitatively and qualitatively outperforms cur rent state of the art in terms of reconstruction quality and novel-view synthesi s.

****
B-Cos Networks: Alignment Is All We Need for Interpretability
Moritz Böhle, Mario Fritz, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10329-10338
We present a new direction for increasing the interpretability of deep neural networks (DNNs) by promoting weight-input alignment during training. For this, we propose to replace the linear transforms in DNNs by our B-cos transform. As we show, a sequence (network) of such transforms induces a single linear transform that faithfully summarises the full model computations. Moreover, the B-cos transform introduces alignment pressure on the weights during optimisation. As a result, those induced linear transforms become highly interpretable and align with task-relevant features. Importantly, the B-cos transform is designed to be compatible with existing architectures and we show that it can easily be integrated into common models such as VGGs, ResNets, InceptionNets, and DenseNets, whilst maintaining similar performance on ImageNet. The resulting explanations are of high visual quality and perform well under quantitative metrics for interpretability. Code available at github.com/moboehle/B-cos.
****
EMOCA: Emotion Driven Monocular Face Capture and Animation
Radek Dan■■ek, Michael J. Black, Timo Bolkart; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20311-20322
As 3D facial avatars become more widely used for communication, it is critical that they faithfully convey emotion. Unfortunately, the best recent methods that regress parametric 3D face models from monocular images are unable to capture the full spectrum of facial expression, such as subtle or extreme emotions. We find the standard reconstruction metrics used for training (landmark reprojection error, photometric error, and face recognition loss) are insufficient to capture high-fidelity expressions. The result is facial geometries that do not match the emotional content of the input image. We address this with EMOCA (EMOtion Capture and Animation), by introducing a novel deep perceptual emotion consistency loss during training, which helps ensure that the reconstructed 3D expression matches the expression depicted in the input image. While EMOCA achieves 3D reconstruction errors that are on par with the current best methods, it significantly outperforms them in terms of the quality of the reconstructed expression and the perceived emotional content. We also directly regress levels of valence and arousal and classify basic expressions from the estimated 3D face parameters. On the task of in-the-wild emotion recognition, our purely geometric approach is on par with the best image-based methods, highlighting the value of 3D geometry in analyzing human behavior. The model and code are publicly available at https://emoca.is.tue.mpg.de.
****
Burst Image Restoration and Enhancement
Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5759-5768
Modern handheld devices can acquire burst image sequence in a quick succession. However, the individual acquired frames suffer from multiple degradations and are misaligned due to camera shake and object motions. The goal of Burst Image Restoration is to effectively combine complimentary cues across multiple burst frames to generate high-quality outputs. Towards this goal, we develop a novel approach by solely focusing on the effective information exchange between burst frames, such that the degradations get filtered out while the actual scene details are preserved and enhanced. Our central idea is to create a set of pseudo-burst features that combine complimentary information from all the input burst frames to seamlessly exchange information. The pseudo-burst representations encode channel-wise features from the original burst images, thus making it easier for the model to learn distinctive information offered by multiple burst frames. However, the pseudo-burst cannot be successfully created unless the individual burst frames are properly aligned to discount inter-frame movements. Therefore, our approach initially extracts preprocessed features from each burst frame and matches th

em using an edge-boosting burst alignment module. The pseudo-burst features are then created and enriched using multi-scale contextual information. Our final st ep is to adaptively aggregate information from the pseudo-burst features to prog ressively increase resolution in multiple stages while merging the pseudo-burst features. In comparison to existing works that usually follow a late fusion sche me with single-stage upsampling, our approach performs favorably, delivering sta te-of-the-art performance on burst super-resolution, burst low-light image enhan cement and burst denoising tasks. Our codes will be publicly released.
****

What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Fact ors

Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, Kevin Smith; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 9225-9234

Transfer learning is a standard technique to transfer knowledge from one domain to another. For applications in medical imaging, transfer from ImageNet has beco me the de-facto approach, despite differences in the tasks and image characteris tics between the domains. However, it is unclear what factors determine whether - and to what extent - transfer learning to the medical domain is useful. The lo ng-standing assumption that features from the source domain get reused has recen tly been called into question. Through a series of experiments on several medica l image benchmark datasets, we explore the relationship between transfer learnin g, data size, the capacity and inductive bias of the model, as well as the dista nce between the source and target domain. Our findings suggest that transfer lea rning is beneficial in most cases, and we characterize the important role featur e reuse plays in its success.
****

Towards Diverse and Natural Scene-Aware 3D Human Motion Synthesis

Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, Bo Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 20460-20469

The ability to synthesize long-term human motion sequences in real-world scenes can facilitate numerous applications. Previous approaches for scene-aware motion synthesis are constrained by pre-defined target objects or positions and thus l imit the diversity of human-scene interactions for synthesized motions. In this paper, we focus on the problem of synthesizing diverse scene-aware human motions under the guidance of target action sequences. To achieve this, we first decomp ose the diversity of scene aware human motions into three aspects, namely intera ction diversity (e.g. sitting on different objects with different poses in the g iven scenes), path diversity (e.g. moving to the target locations following diff erent paths), and the motion diversity (e.g. having various body movements durin g moving). Based on this factorized scheme, a hierarchical framework is proposed with each sub-module responsible for modeling one aspect. We assess the effecti veness of our framework on two challenging datasets for scene-aware human motion synthesis. The experiment results show that the proposed framework remarkably o utperforms the previous methods in terms of diversity and naturalness.
****

Quarantine: Sparsity Can Uncover the Trojan Attack Trigger for Free

Tianlong Chen, Zhenyu Zhang, Yihua Zhang, Shiyu Chang, Sijia Liu, Zhangyang Wang ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 598-609

Trojan attacks threaten deep neural networks (DNNs) by poisoning them to behave normally on most samples, yet to produce manipulated results for inputs attached with a particular trigger. Several works attempt to detect whether a given DNN has been injected with a specific trigger during the training. In a parallel lin e of research, the lottery ticket hypothesis reveals the existence of sparse sub networks which are capable of reaching competitive performance as the dense netw ork after independent training. Connecting these two dots, we investigate the pr oblem of Trojan DNN detection from the brand new lens of sparsity, even when no clean training data is available. Our crucial observation is that the Trojan fea

tures are significantly more stable to network pruning than benign features. Leveraging that, we propose a novel Trojan network detection regime: first locating a "winning Trojan lottery ticket" which preserves nearly full Trojan information yet only chance-level performance on clean inputs; then recovering the trigger embedded in this already isolated subnetwork. Extensive experiments on various datasets, i.e., CIFAR-10, CIFAR-100, and ImageNet, with different network architectures, i.e., VGG-16, ResNet-18, ResNet-20s, and DenseNet-100 demonstrate the effectiveness of our proposal. Codes are available at https://github.com/VITA-Group/Backdoor-LTH.

****

Audio-Visual Speech Codecs: Rethinking Audio-Visual Speech Enhancement by Re-Synthesis

Karren Yang, Dejan Markovi■, Steven Krenn, Vasu Agrawal, Alexander Richard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8227-8237

Since facial actions such as lip movements contain significant information about speech content, it is not surprising that audio-visual speech enhancement methods are more accurate than their audio-only counterparts. Yet, state-of-the-art approaches still struggle to generate clean, realistic speech without noise artifacts and unnatural distortions in challenging acoustic environments. In this paper, we propose a novel audio-visual speech enhancement framework for high-fidelity telecommunications in AR/VR. Our approach leverages audio-visual speech cues to generate the codes of a neural speech codec, enabling efficient synthesis of clean, realistic speech from noisy signals. Given the importance of speaker-specific cues in speech, we focus on developing personalized models that work well for individual speakers. We demonstrate the efficacy of our approach on a new audio-visual speech dataset collected in an unconstrained, large vocabulary setting, as well as existing audio-visual datasets, outperforming speech enhancement baselines on both quantitative metrics and human evaluation studies. Please see the supplemental video for qualitative results.

****

Localized Adversarial Domain Generalization

Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, Adam P. Harrison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7108-7118

Deep learning methods can struggle to handle domain shifts not seen in training data, which can cause them to not generalize well to unseen domains. This has led to research attention on domain generalization (DG), which aims to the model's generalization ability to out-of-distribution. Adversarial domain generalization is a popular approach to DG, but conventional approaches (1) struggle to sufficiently align features so that local neighborhoods are mixed across domains; and (2) can suffer from feature space over collapse which can threaten generalization performance. To address these limitations, we propose localized adversarial domain generalization with space compactness maintenance (LADG) which constitutes two major contributions. First, we propose an adversarial localized classifier as the domain discriminator, along with a principled primary branch. This constructs a min-max game whereby the aim of the featurizer is to produce locally mixed domains. Second, we propose to use a coding-rate loss to alleviate feature space over collapse. We conduct comprehensive experiments on the Wilds DG benchmark to validate our approach, where LADG outperforms leading competitors on most datasets.

****

X-Trans2Cap: Cross-Modal Knowledge Transfer Using Transformer for 3D Dense Captioning

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, Zhen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8563-8573

3D dense captioning aims to describe individual objects by natural language in 3D scenes, where 3D scenes are usually represented as RGB-D scans or point clouds. However, only exploiting single modal information, e.g., point cloud, previous

approaches fail to produce faithful descriptions. Though aggregating 2D features into point clouds may be beneficial, it introduces an extra computational burden, especially in inference phases. In this study, we investigate a cross-modal knowledge transfer using Transformer for 3D dense captioning, X-Trans2Cap, to effectively boost the performance of single-modal 3D caption through knowledge distillation using a teacher-student framework. In practice, during the training phase, the teacher network exploits auxiliary 2D modality and guides the student network that only takes point clouds as input through the feature consistency constraints. Owing to the well-designed cross-modal feature fusion module and the feature alignment in the training phase, X-Trans2Cap acquires rich appearance information embedded in 2D images with ease. Thus, a more faithful caption can be generated only using point clouds during the inference. Qualitative and quantitative results confirm that X-Trans2Cap outperforms previous state-of-the-art by a large margin, i.e., about +21 and about +16 absolute CIDEr score on ScanRefer and Nr3D datasets, respectively.

****

How Much Does Input Data Type Impact Final Face Model Accuracy?

Jiahao Luo, Fahim Hasan Khan, Issei Mori, Akila de Silva, Eric Sandoval Ruezga, Minghao Liu, Alex Pang, James Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18985-18994

Face models are widely used in image processing and other domains. The input data to create a 3D face model ranges from accurate laser scans to simple 2D RGB photographs. These input data types are typically deficient either due to missing regions, or because they are under-constrained. As a result, reconstruction methods include embedded priors encoding the valid domain of faces. System designers must choose a source of input data and then choose a reconstruction method to obtain a usable 3D face. If a particular application domain requires accuracy X, which kinds of input data are suitable? Does the input data need to be 3D, or will 2D data suffice? This paper takes a step toward answering these questions using synthetic data. A ground truth dataset is used to analyze accuracy obtainable from 2D landmarks, 3D landmarks, low quality 3D, high quality 3D, texture color, normals, dense 2D image data, and when regions of the face are missing. Since the data is synthetic it can be analyzed both with and without measurement error. This idealized synthetic analysis is then compared to real results from several methods for constructing 3D faces from 2D photographs. The experimental results suggest that accuracy is severely limited when only 2D raw input data exists.

****

Image-to-Lidar Self-Supervised Distillation for Autonomous Driving Data

Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, Renaud Marlet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9891-9901

Segmenting or detecting objects in sparse Lidar point clouds are two important tasks in autonomous driving to allow a vehicle to act safely in its 3D environment. The best performing methods in 3D semantic segmentation or object detection rely on a large amount of annotated data. Yet annotating 3D Lidar data for these tasks is tedious and costly. In this context, we propose a self-supervised pre-training method for 3D perception models that is tailored to autonomous driving data. Specifically, we leverage the availability of synchronized and calibrated image and LiDAR sensors in autonomous driving setups for distilling self-supervised pre-trained image representations into 3D models. Hence, our method does not require any point cloud nor image annotations. The key ingredient of our method is the use of superpixels which are used to pool 3D point features and 2D pixel features in visually similar regions. We then train a 3D network on the self-supervised task of matching these pooled point features with the corresponding pooled image pixel features. The advantages of contrasting regions obtained by superpixels are that: (1) grouping together pixels and points of visually coherent regions leads to a more meaningful contrastive task that produces features well adapted to 3D semantic segmentation and 3D object detection; (2) all the different regions have the same weight in the contrastive loss regardless of the number of 3D points sampled in these regions; (3) it mitigates the noise produced by inc

orrect matching of points and pixels due to occlusions between the different sen
sors. Extensive experiments on autonomous driving datasets demonstrate the abili
ty of our image-to-Lidar distillation strategy to produce 3D representations tha
t transfer well on semantic segmentation and object detection tasks.
****

HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video
Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, Ira Keme
lmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2022, pp. 16210-16220
We introduce a free-viewpoint rendering method -- HumanNeRF -- that works on a g
iven monocular video of a human performing complex body motions, e.g. a video fr
om YouTube. Our method enables pausing the video at any frame and rendering the
subject from arbitrary new camera viewpoints or even a full 360-degree camera pa
th for that particular frame and body pose. This task is particularly challengin
g, as it requires synthesizing photorealistic details of the body, as seen from
various camera angles that may not exist in the input video, as well as synthesi
zing fine details such as cloth folds and facial appearance. Our method optimize
s for a volumetric representation of the person in a canonical T-pose, in concer
t with a motion field that maps the estimated canonical representation to every
frame of the video via backward warps. The motion field is decomposed into skele
tal rigid and non-rigid motions, produced by deep networks. We show significant
performance improvements over prior work, and compelling examples of free-viewpo
int renderings from monocular video of moving humans in challenging uncontrolled
 capture scenarios.
****

PoseKernelLifter: Metric Lifting of 3D Human Pose Using Sound
Zhijian Yang, Xiaoran Fan, Volkan Isler, Hyun Soo Park; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1317
9-13189
Reconstructing the 3D pose of a person in metric scale from a single view image
is a geometrically ill-posed problem. For example, we can not measure the exact
distance of a person to the camera from a single view image without additional s
cene assumptions (e.g., known height). Existing learning based approaches circum
vent this issue by reconstructing the 3D pose up to scale. However, there are ma
ny applications such as virtual telepresence, robotics, and augmented reality th
at require metric scale reconstruction. In this paper, we show that audio signal
s recorded along with an image, provide complementary information to reconstruct
 the metric 3D pose of the person. The key insight is that as the audio signals
traverse across the 3D space, their interactions with the body provide metric in
formation about the body's pose. Based on this insight, we introduce a time-inva
riant transfer function called pose kernel---the impulse response of audio signa
ls induced by the body pose. The main properties of the pose kernel are that (1)
 its envelope highly correlates with 3D pose, (2) the time response corresponds
to arrival time, indicating the metric distance to the microphone, and (3) it is
 invariant to changes in the scene geometry configurations. Therefore, it is rea
dily generalizable to unseen scenes. We design a multi-stage 3D CNN that fuses a
udio and visual signals and learns to reconstruct 3D pose in a metric scale. We
show that our multi-modal method produces accurate metric reconstruction in real
 world scenes, which is not possible with state-of-the-art lifting approaches in
cluding parametric mesh regression and depth regression.
****

Which Images To Label for Few-Shot Medical Landmark Detection?
Quan Quan, Qingsong Yao, Jun Li, S. Kevin Zhou; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20606-20616
The success of deep learning methods relies on the availability of well-labeled
large-scale datasets. However, for medical images, annotating such abundant trai
ning data often requires experienced radiologists and consumes their limited tim
e. Few-shot learning is developed to alleviate this burden, which achieves compe
titive performance with only several labeled data. However, a crucial yet previo
usly overlooked problem in few-shot learning is about the selection of the templ

ate images for annotation before learning, which affects the final performance. We herein propose a novel Sample Choosing Policy (SCP) to select "the most worthy" images as the templates, in the context of medical landmark detection. SCP consists of three parts: 1) Self-supervised training for building a pre-trained deep model to extract features from radiological images, 2) Key Point Proposal for localizing informative patches, and 3) Representative Score Estimation for searching most representative samples or templates. The performance of SCP is demonstrated by various experiments on several widely-used public datasets. For one-shot medical landmark detection, the mean radial errors on Cephalometric and HandXray datasets are reduced from 3.595mm to 3.083mm and 4.114mm to 2.653mm, respectively.

****

Why Discard if You Can Recycle?: A Recycling Max Pooling Module for 3D Point Cloud Analysis

Jiajing Chen, Burak Kakillioglu, Huantao Ren, Senem Velipasalar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 559-567

In recent years, most 3D point cloud analysis models have focused on developing either new network architectures or more efficient modules for aggregating point features from a local neighborhood. Regardless of the network architecture or the methodology used for improved feature learning, these models share one thing, which is the use of max-pooling in the end to obtain permutation invariant features. We first show that this traditional approach causes only a fraction of 3D points contribute to the permutation invariant features, and discards the rest of the points. In order to address this issue and improve the performance of any baseline 3D point classification or segmentation model, we propose a new module, referred to as the Recycling MaxPooling (RMP) module, to recycle and utilize the features of some of the discarded points. We incorporate a refinement loss that uses the recycled features to refine the prediction loss obtained from the features kept by traditional max-pooling. To the best of our knowledge, this is the first work that explores recycling of still useful points that are traditionally discarded by max-pooling. We demonstrate the effectiveness of the proposed RMP module by incorporating it into several milestone baselines and state-of-the-art networks for point cloud classification and indoor semantic segmentation tasks. We show that RPM, without any bells and whistles, consistently improves the performance of all the tested networks by using the same base network implementation and hyper-parameters. The code is provided in the supplementary material.

****

Explaining Deep Convolutional Neural Networks via Latent Visual-Semantic Filter Attention

Yu Yang, Seungbae Kim, Jungseock Joo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8333-8343

Interpretability is an important property for visual models as it helps researchers and users understand the internal mechanism of a complex model. However, generating semantic explanations about the learned representation is challenging without direct supervision to produce such explanations. We propose a general framework, Latent Visual Semantic Explainer (LaViSE), to teach any existing convolutional neural network to generate text descriptions about its own latent representations at the filter level. Our method constructs a mapping between the visual and semantic spaces using generic image datasets, using images and category names. It then transfers the mapping to the target domain which does not have semantic labels. The proposed framework employs a modular structure and enables to analyze any trained network whether or not its original training data is available. We show that our method can generate novel descriptions for learned filters beyond the set of categories defined in the training dataset and perform an extensive evaluation on multiple datasets. We also demonstrate a novel application of our method for unsupervised dataset bias analysis which allows us to automatically discover hidden biases in datasets or compare different subsets without using additional labels. The dataset and code are made public to facilitate further research.

****
AlignQ: Alignment Quantization With ADMM-Based Correlation Preservation
Ting-An Chen, De-Nian Yang, Ming-Syan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12538-12547
Quantization is an efficient network compression approach to reduce the inference time. However, existing approaches ignored the distribution difference between training and testing data, thereby inducing a large quantization error in inference. To address this issue, we propose a new quantization scheme, Alignment Quantization with ADMM-based Correlation Preservation (AlignQ), which exploits the cumulative distribution function (CDF) to align the data to be i.i.d. (independently and identically distributed) for quantization error minimization. Afterward, our theoretical analysis indicates that the significant changes in data correlations after the quantization induce a large quantization error. Accordingly, we aim to preserve the relationship of data from the original space to the aligned quantization space for retaining the prediction information. We design an optimization process by leveraging the Alternating Direction Method of Multipliers (ADMM) optimization to minimize the differences in data correlations before and after the alignment and quantization. In experiments, we visualize non-i.i.d. in training and testing data in the benchmark. We further adopt domain shift data to compare AlignQ with the state-of-the-art. Experimental results show that AlignQ achieves significant performance improvements, especially in low-bit models. Code is available at https://github.com/tinganchen/AlignQ.git.
****
Self-Distillation From the Last Mini-Batch for Consistency Regularization
Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, Yandong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11943-11952
Knowledge distillation (KD) shows a bright promise as a powerful regularization strategy to boost generalization ability by leveraging learned sample-level soft targets. Yet, employing a complex pre-trained teacher network or an ensemble of peer students in existing KD is both time-consuming and computationally costly. Various self KD methods have been proposed to achieve higher distillation efficiency. However, they either require extra network architecture modification, or are difficult to parallelize. To cope with these challenges, we propose an efficient and reliable self-distillation framework, named Self-Distillation from Last Mini-Batch (DLB). Specifically, we rearrange the sequential sampling by constraining half of each mini-batch coinciding with the previous iteration. Meanwhile, the rest half will coincide with the upcoming iteration. Afterwards, the former half mini-batch distills on-the-fly soft targets generated in the previous iteration. Our proposed mechanism guides the training stability and consistency, resulting in robustness to label noise. Moreover, our method is easy to implement, without taking up extra run-time memory or requiring model structure modification. Experimental results on three classification benchmarks illustrate that our approach can consistently outperform state-of-the-art self-distillation approaches with different network architectures. Additionally, our method shows strong compatibility with augmentation strategies by gaining additional performance improvement.
****
Interactive Multi-Class Tiny-Object Detection
Chunggi Lee, Seonwook Park, Heon Song, Jeongun Ryu, Sanghoon Kim, Haejoon Kim, Sérgio Pereira, Donggeun Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14136-14145
Annotating tens or hundreds of tiny objects in a given image is laborious yet crucial for a multitude of Computer Vision tasks. Such imagery typically contains objects from various categories, yet the multi-class interactive annotation setting for the detection task has thus far been unexplored. To address these needs, we propose a novel interactive annotation method for multiple instances of tiny objects from multiple classes, based on a few point-based user inputs. Our approach, C3Det, relates the full image context with annotator inputs in a local and global manner via late-fusion and feature-correlation, respectively. We perform

experiments on the Tiny-DOTA and LCell datasets using both two-stage and one-stage object detection architectures to verify the efficacy of our approach. Our approach outperforms existing approaches in interactive annotation, achieving higher mAP with fewer clicks. Furthermore, we validate the annotation efficiency of our approach in a user study where it is shown to be 2.85x faster and yield only 0.36x task load (NASA-TLX, lower is better) compared to manual annotation. The code is available at https://github.com/ChungYi347/Interactive-Multi-Class-Tiny-Object-Detection.

****

Learning From Pixel-Level Noisy Label: A New Perspective for Light Field Saliency Detection

Mingtao Feng, Kendong Liu, Liang Zhang, Hongshan Yu, Yaonan Wang, Ajmal Mian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1756-1766

Saliency detection with light field images is becoming attractive given the abundant cues available, however, this comes at the expense of large-scale pixel level annotated data which is expensive to generate. In this paper, we propose to learn light field saliency from pixel-level noisy labels obtained from unsupervised hand crafted featured-based saliency methods. Given this goal, a natural question is: can we efficiently incorporate the relationships among light field cues while identifying clean labels in a unified framework? We address this question by formulating the learning as a joint optimization of intra light field features fusion stream and inter scenes correlation stream to generate the predictions. Specially, we first introduce a pixel forgetting guided fusion module to mutually enhance the light field features and exploit pixel consistency across iterations to identify noisy pixels. Next, we introduce a cross scene noise penalty loss for better reflecting latent structures of training data and enabling the learning to be invariant to noise. Extensive experiments on multiple benchmark datasets demonstrate the superiority of our framework showing that it learns saliency prediction comparable to state-of-the-art fully supervised light field saliency methods. Our code is available at https://github.com/OLobbCode/NoiseLF.

****

UBoCo: Unsupervised Boundary Contrastive Learning for Generic Event Boundary Detection

Hyolim Kang, Jinwoo Kim, Taehyun Kim, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20073-20082

Generic Event Boundary Detection (GEBD) is a newly suggested video understanding task that aims to find one level deeper semantic boundaries of events. Bridging the gap between natural human perception and video understanding, it has various potential applications, including interpretable and semantically valid video parsing. Still at an early development stage, existing GEBD solvers are simple extensions of relevant video understanding tasks, disregarding GEBD's distinctive characteristics. In this paper, we propose a novel framework for unsupervised/supervised GEBD, by using the Temporal Self-similarity Matrix (TSM) as the video representation. The new Recursive TSM Parsing (RTP) algorithm exploits local diagonal patterns in TSM to detect boundaries, and it is combined with the Boundary Contrastive (BoCo) loss to train our encoder to generate more informative TSMs. Our framework can be applied to both unsupervised and supervised settings, with both achieving state-of-the-art performance by a huge margin in GEBD benchmark. Especially, our unsupervised method outperforms previous state-of-the-art "supervised" model, implying its exceptional efficacy.

****

Multi-View Depth Estimation by Fusing Single-View Depth Probability With Multi-View Geometry

Gwangbin Bae, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2842-2851

Multi-view depth estimation methods typically require the computation of a multi-view cost-volume, which leads to huge memory consumption and slow inference. Furthermore, multi-view matching can fail for texture-less surfaces, reflective su

rfaces and moving objects. For such failure modes, single-view depth estimation methods are often more reliable. To this end, we propose MaGNet, a novel framework for fusing single-view depth probability with multi-view geometry, to improve the accuracy, robustness and efficiency of multi-view depth estimation. For each frame, MaGNet estimates a single-view depth probability distribution, parameterized as a pixel-wise Gaussian. The distribution estimated for the reference frame is then used to sample per-pixel depth candidates. Such probabilistic sampling enables the network to achieve higher accuracy while evaluating fewer depth candidates. We also propose depth consistency weighting for the multi-view matching score, to ensure that the multi-view depth is consistent with the single-view predictions. The proposed method achieves state-of-the-art performance on ScanNet, 7-Scenes and KITTI. Qualitative evaluation demonstrates that our method is more robust against challenging artifacts such as texture-less/reflective surfaces and moving objects. Our code and model weights are available at https://github.com/baegwangbin/MaGNet.

****

Learning To Collaborate in Decentralized Learning of Personalized Models

Shuangtong Li, Tianyi Zhou, Xinmei Tian, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9766-9775

Learning personalized models for user-customized computer-vision tasks is challenging due to the limited private-data and computation available on each edge device. Decentralized learning (DL) can exploit the images distributed over devices on a network topology to train a global model but is not designed to train personalized models for different tasks or optimize the topology. Moreover, the mixing weights used to aggregate neighbors' gradient messages in DL can be sub-optimal for personalization since they are not adaptive to different nodes/tasks and learning stages. In this paper, we dynamically update the mixing-weights to improve the personalized model for each node's task and meanwhile learn a sparse topology to reduce communication costs. Our first approach, "learning to collaborate (L2C)", directly optimizes the mixing weights to minimize the local validation loss per node for a pre-defined set of nodes/tasks. In order to produce mixing weights for new nodes or tasks, we further develop "meta-L2C", which learns an attention mechanism to automatically assign mixing weights by comparing two nodes' model updates. We evaluate both methods on diverse benchmarks and experimental settings for image classification. Thorough comparisons to both classical and recent methods for IID/non-IID decentralized and federated learning demonstrate our method's advantages in identifying collaborators among nodes, learning sparse topology, and producing better personalized models with low communication and computational cost.

****

CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, Jing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3835-3844

We present CLIP-NeRF, a multi-modal 3D object manipulation method for neural radiance fields (NeRF). By leveraging the joint language-image embedding space of the recent Contrastive Language-Image Pre-Training (CLIP) model, we propose a unified framework that allows manipulating NeRF in a user-friendly way, using either a short text prompt or an exemplar image. Specifically, to combine the novel view synthesis capability of NeRF and the controllable manipulation ability of latent representations from generative models, we introduce a disentangled conditional NeRF architecture that allows individual control over both shape and appearance. This is achieved by performing the shape conditioning via applying a learned deformation field to the positional encoding and deferring color conditioning to the volumetric rendering stage. To bridge this disentangled latent representation to the CLIP embedding, we design two code mappers that take a CLIP embedding as input and update the latent codes to reflect the targeted editing. The mappers are trained with a CLIP-based matching loss to ensure the manipulation accuracy. Furthermore, we propose an inverse optimization method that accurately pro

jects an input image to the latent codes for manipulation to enable editing on r
eal images. We evaluate our approach by extensive experiments on a variety of te
xt prompts and exemplar images and also provide an intuitive editing interface f
or real-time user interaction.
****

ART-Point: Improving Rotation Robustness of Point Cloud Classifiers via Adversar
ial Rotation
Ruibin Wang, Yibo Yang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2022, pp. 14371-14380
Point cloud classifiers with rotation robustness have been widely discussed in t
he 3D deep learning community. Most proposed methods either use rotation invaria
nt descriptors as inputs or try to design rotation equivariant networks. However
, robust models generated by these methods have limited performance under clean
aligned datasets due to modifications on the original classifiers or input space
. In this study, for the first time, we show that the rotation robustness of poi
nt cloud classifiers can also be acquired via adversarial training with better p
erformance on both rotated and clean datasets. Specifically, our proposed framew
ork named ART-Point regards the rotation of the point cloud as an attack and imp
roves rotation robustness by training the classifier on inputs with Adversarial
RoTations. We contribute an axis-wise rotation attack that uses back-propagated
gradients of the pre-trained model to effectively find the adversarial rotations
. To avoid model over-fitting on adversarial inputs, we construct rotation pools
 that leverage the transferability of adversarial rotations among samples to inc
rease the diversity of training data. Moreover, we propose a fast one-step optim
ization to efficiently reach the final robust model. Experiments show that our p
roposed rotation attack achieves a high success rate and ART-Point can be used o
n most existing classifiers to improve the rotation robustness while showing bet
ter performance on clean datasets than state-of-the-art methods.
****

Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields
Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, Prat
ul P. Srinivasan; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2022, pp. 5491-5500
Neural Radiance Fields (NeRF) is a popular view synthesis technique that represe
nts a scene as a continuous volumetric function, parameterized by multilayer per
ceptrons that provide the volume density and view-dependent emitted radiance at
each location. While NeRF-based techniques excel at representing fine geometric
structures with smoothly varying view-dependent appearance, they often fail to a
ccurately capture and reproduce the appearance of glossy surfaces. We address th
is limitation by introducing Ref-NeRF, which replaces NeRF's parameterization of
 view-dependent outgoing radiance with a representation of reflected radiance an
d structures this function using a collection of spatially-varying scene propert
ies. We show that together with a regularizer on normal vectors, our model signi
ficantly improves the realism and accuracy of specular reflections. Furthermore,
 we show that our model's internal representation of outgoing radiance is interp
retable and useful for scene editing.
****

360-Attack: Distortion-Aware Perturbations From Perspective-Views
Yunjian Zhang, Yanwei Liu, Jinxia Liu, Jingbo Miao, Antonios Argyriou, Liming Wa
ng, Zhen Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2022, pp. 15035-15044
The application of deep neural networks (DNNs) on 360-degree images has achieved
 remarkable progress in the recent years. However, DNNs have been demonstrated t
o be vulnerable to well-crafted adversarial examples, which may trigger severe s
afety problems in the real-world applications based on 360-degree images. In thi
s paper, we propose an adversarial attack targeting spherical images, called 360
-attactk, that transfers adversarial perturbations from perspective-view (PV) im
ages to a final adversarial spherical image. Given a target spherical image, we
first represent it with a set of planar PV images, and then perform 2D attacks o
n them to obtain adversarial PV images. Considering the issue of the projective

distortion between spherical and PV images, we propose a distortion-aware attack to reduce the negative impact of distortion on attack. Moreover, to reconstruct the final adversarial spherical image with high aggressiveness, we calculate the spherical saliency map with a novel spherical spectrum method and next propose a saliency-aware fusion strategy that merges multiple inverse perspective projections for the same position on the spherical image. Extensive experimental results show that 360-attack is effective for disturbing spherical images in the black-box setting. Our attack also proves the presence of adversarial transferability from Z2 to SO3 groups.

****

Targeted Supervised Contrastive Learning for Long-Tailed Recognition

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S. Feris, Piotr Indyk, Dina Katabi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6918-6928

Real-world data often exhibits long tail distributions with heavy class imbalance, where the majority classes can dominate the training process and alter the decision boundaries of the minority classes. Recently, researchers have investigated the potential of supervised contrastive learning for long-tailed recognition, and demonstrated that it provides a strong performance gain. In this paper, we show that while supervised contrastive learning can help improve performance, past baselines suffer from poor uniformity brought in by imbalanced data distribution. This poor uniformity manifests in samples from the minority class having poor separability in the feature space. To address this problem, we propose targeted supervised contrastive learning (TSC), which improves the uniformity of the feature distribution on the hypersphere. TSC first generates a set of targets uniformly distributed on a hypersphere. It then makes the features of different classes converge to these distinct and uniformly distributed targets during training. This forces all classes, including minority classes, to maintain a uniform distribution in the feature space, improves class boundaries, and provides better generalization even in the presence of long-tail data. Experiments on multiple datasets show that TSC achieves state-of-the-art performance on long-tailed recognition tasks.

****

Both Style and Fog Matter: Cumulative Domain Adaptation for Semantic Foggy Scene Understanding

Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, Chia-Wen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18922-18931

Although considerable progress has been made in semantic scene understanding under clear weather, it is still a tough problem under adverse weather conditions, such as dense fog, due to the uncertainty caused by imperfect observations. Besides, difficulties in collecting and labeling foggy images hinder the progress of this field. Considering the success in semantic scene understanding under clear weather, we think it is reasonable to transfer knowledge learned from clear images to the foggy domain. As such, the problem becomes to bridge the domain gap between clear images and foggy images. Unlike previous methods that mainly focus on closing the domain gap caused by fog --- defogging the foggy images or fogging the clear images, we propose to alleviate the domain gap by considering fog influence and style variation simultaneously. The motivation is based on our finding that the style-related gap and the fog-related gap can be divided and closed respectively, by adding an intermediate domain. Thus, we propose a new pipeline to cumulatively adapt style, fog and the dual-factor (style and fog). Specifically, we devise a unified framework to disentangle the style factor and the fog factor separately, and then the dual-factor from images in different domains. Furthermore, we collaborate the disentanglement of three factors with a novel cumulative loss to thoroughly disentangle these three factors. Our method achieves the state-of-the-art performance on three benchmarks and shows generalization ability in rainy and snowy scenes.

****

Ev-TTA: Test-Time Adaptation for Event-Based Object Recognition

Junho Kim, Inwoo Hwang, Young Min Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17745-17754
We introduce Ev-TTA, a simple, effective test-time adaptation algorithm for event-based object recognition. While event cameras are proposed to provide measurements of scenes with fast motions or drastic illumination changes, many existing event-based recognition algorithms suffer from performance deterioration under extreme conditions due to significant domain shifts. Ev-TTA mitigates the severe domain gaps by fine-tuning the pre-trained classifiers during the test phase using loss functions inspired by the spatio-temporal characteristics of events. Since the event data is a temporal stream of measurements, our loss function enforces similar predictions for adjacent events to quickly adapt to the changed environment online. Also, we utilize the spatial correlations between two polarities of events to handle noise under extreme illumination, where different polarities of events exhibit distinctive noise distributions. Ev-TTA demonstrates a large amount of performance gain on a wide range of event-based object recognition tasks without extensive additional training. Our formulation can be successfully applied regardless of input representations and further extended into regression tasks. We expect Ev-TTA to provide the key technique to deploy event-based vision algorithms in challenging real-world applications where significant domain shift is inevitable.
****

Balanced Contrastive Learning for Long-Tailed Visual Recognition
Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6908-6917
Real-world data typically follow a long-tailed distribution, where a few majority categories occupy most of the data while most minority categories contain a limited number of samples. Classification models minimizing cross-entropy struggle to represent and classify the tail classes. Although the problem of learning unbiased classifiers has been well studied, methods for representing imbalanced data are under-explored. In this paper, we focus on representation learning for imbalanced data. Recently, supervised contrastive learning has shown promising performance on balanced data recently. However, through our theoretical analysis, we find that for long-tailed data, it fails to form a regular simplex which is an ideal geometric configuration for representation learning. To correct the optimization behavior of SCL and further improve the performance of long-tailed visual recognition, we propose a novel loss for balanced contrastive learning (BCL). Compared with SCL, we have two improvements in BCL: class-averaging, which balances the gradient contribution of negative classes; class-complement, which allows all classes to appear in every mini-batch. The proposed balanced contrastive learning (BCL) method satisfies the condition of forming a regular simplex and assists the optimization of cross-entropy. Equipped with BCL, the proposed two-branch framework can obtain a stronger feature representation and achieve competitive performance on long-tailed benchmark datasets such as CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist2018.
****

Slimmable Domain Adaptation
Rang Meng, Weijie Chen, Shicai Yang, Jie Song, Luojun Lin, Di Xie, Shiliang Pu, Xinchao Wang, Mingli Song, Yueting Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7141-7150
Vanilla unsupervised domain adaptation methods tend to optimize the model with fixed neural architecture, which is not very practical in real-world scenarios since the target data is usually processed by different resource-limited devices. It is therefore of great necessity to facilitate architecture adaptation across various devices. In this paper, we introduce a simple framework, Slimmable Domain Adaptation, to improve cross-domain generalization with a weight-sharing model bank, from which models of different capacities can be sampled to accommodate different accuracy-efficiency trade-offs. The main challenge in this framework lies in simultaneously boosting the adaptation performance of numerous models in the model bank. To tackle this problem, we develop a Stochastic EnsEmble Distilla

tion method to fully exploit the complementary knowledge in the model bank for i nter-model interaction. Nevertheless, considering the optimization conflict betw een inter-model interaction and intra-model adaptation, we augment the existing bi-classifier domain confusion architecture into an Optimization-Separated Tri-C lassifier counterpart. After optimizing the model bank, architecture adaptation is leveraged via our proposed Unsupervised Performance Evaluation Metric. Under various resource constraints, our framework surpasses other competing approaches by a very large margin on multiple benchmarks. It is also worth emphasizing tha t our framework can preserve the performance improvement against the source-only model even when the computing complexity is reduced to 1/64. Code will be avail able at https://github.com/HIK-LAB/SlimDA.

****

Bandits for Structure Perturbation-Based Black-Box Attacks To Graph Neural Netwo rks With Theoretical Guarantees

Binghui Wang, Youqi Li, Pan Zhou; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2022, pp. 13379-13387

Graph neural networks (GNNs) have achieved state-ofthe-art performance in many g raph-based tasks such as node classification and graph classification. However, many recent works have demonstrated that an attacker can mislead GNN models by s lightly perturbing the graph structure. Existing attacks to GNNs are either unde r the less practical threat model where the attacker is assumed to access the GN N model parameters, or under the practical black-box threat model but consider p erturbing node features that are shown to be not enough effective. In this paper , we aim to bridge this gap and consider black-box attacks to GNNs with structur e perturbation as well as with theoretical guarantees. We propose to address thi s challenge through bandit techniques. Specifically, we formulate our attack as an online optimization with bandit feedback. This original problem is essentiall y NP-hard due to the fact that perturbing the graph structure is a binary optimi zation problem. We then propose an online attack based on bandit optimization wh ich is proven to be sublinear to the query number T, i.e., O(N^ 1/2 T^ 3/4) whe re N is the number of nodes in the graph. Finally, we evaluate our proposed atta ck by conducting experiments over multiple datasets and GNN models. The experime ntal results on various citation graphs and image graphs show that our attack is both effective and efficient.

****

NODEO: A Neural Ordinary Differential Equation Based Optimization Framework for Deformable Image Registration

Yifan Wu, Tom Z. Jiahao, Jiancong Wang, Paul A. Yushkevich, M. Ani Hsieh, James C. Gee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 20804-20813

Deformable image registration (DIR), aiming to find spatial correspondence betwe en images, is one of the most critical problems in the domain of medical image a nalysis. In this paper, we present a novel, generic, and accurate diffeomorphic image registration framework that utilizes neural ordinary differential equation s (NODEs). We model each voxel as a moving particle and consider the set of all voxels in a 3D image as a high-dimensional dynamical system whose trajectory det ermines the targeted deformation field. Our method leverages deep neural network s for their expressive power in modeling dynamical systems, and simultaneously o ptimizes for a dynamical system between the image pairs and the corresponding tr ansformation. Our formulation allows various constraints to be imposed along the transformation to maintain desired regularities. Our experiment results show th at our method outperforms the benchmarks under various metrics. Additionally, we demonstrate the feasibility to expand our framework to register multiple image sets using a unified form of transformation, which could possibly serve a wider range of applications.

****

DIP: Deep Inverse Patchmatch for High-Resolution Optical Flow

Zihua Zheng, Ni Nie, Zhi Ling, Pengfei Xiong, Jiangyu Liu, Hao Wang, Jiankun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2022, pp. 8925-8934

Recently, the dense correlation volume method achieves state-of-the-art performa
nce in optical flow. However, the correlation volume computation requires a lot
of memory, which makes prediction difficult on high-resolution images. In this p
aper, we propose a novel Patchmatch-based framework to work on high-resolution o
ptical flow estimation. Specifically, we introduce the first end-to-end Patchmat
ch based deep learning optical flow. It can get high-precision results with lowe
r memory benefiting from propagation and local search of Patchmatch. Furthermore
, a new inverse propagation is proposed to decouple the complex operations of pr
opagation, which can significantly reduce calculations in multiple iterations. A
t the time of submission, our method ranks first on all the metrics on the popul
ar KITTI2015 benchmark, and ranks second on EPE on the Sintel clean benchmark am
ong published optical flow methods. Experiment shows our method has a strong cro
ss-dataset generalization ability that the F1-all achieves 13.73%, reducing 21%
from the best published result 17.4% on KITTI2015. What's more, our method shows
 a good details preserving result on the high-resolution dataset DAVIS and consu
mes 2x less memory than RAFT.
****

Few-Shot Object Detection With Fully Cross-Transformer
Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, Shih-Fu Chang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
2, pp. 5321-5330
Few-shot object detection (FSOD), with the aim to detect novel objects using ver
y few training examples, has recently attracted great research interest in the c
ommunity. Metric-learning based methods have been demonstrated to be effective f
or this task using a two-branch based siamese network, and calculate the similar
ity between image regions and few-shot examples for detection. However, in previ
ous works, the interaction between the two branches is only restricted in the de
tection head, while leaving the remaining hundreds of layers for separate featur
e extraction. Inspired by the recent work on vision transformers and vision-lang
uage transformers, we propose a novel Fully Cross-Transformer based model (FCT)
for FSOD by incorporating cross-transformer into both the feature backbone and d
etection head. The asymmetric-batched cross-attention is proposed to aggregate t
he key information from the two branches with different batch sizes. Our model c
an improve the few-shot similarity learning between the two branches by introduc
ing the multi-level interactions. Comprehensive experiments on both PASCAL VOC a
nd MSCOCO FSOD benchmarks demonstrate the effectiveness of our model.
****

Pyramid Architecture for Multi-Scale Processing in Point Cloud Segmentation
Dong Nie, Rui Lan, Ling Wang, Xiaofeng Ren; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17284-17294
Semantic segmentation of point cloud data is a critical task for autonomous driv
ing and other applications. Recent advances of point cloud segmentation are main
ly driven by new designs of local aggregation operators and point sampling metho
ds. Unlike image segmentation, few efforts have been made to understand the fund
amental issue of scale and how scales should interact and be fused. In this work
, we investigate how to efficiently and effectively integrate features at varyin
g scales and varying stages in a point cloud segmentation network. In particular
, we open up the commonly used encoder-decoder architecture, and design scale py
ramid architectures that allow information to flow more freely and systematicall
y, both laterally and upward/downward in scale. Moreover, a cross-scale attentio
n feature learning block has been designed to enhance the multi-scale feature fu
sion which occurs everywhere in the network. Such a design of multi-scale proces
sing and fusion gains large improvements in accuracy without adding much additio
nal computation. When built on top of the popular KPConv network, we see consist
ent improvements on a wide range of datasets, including achieving state-of-the-a
rt performance on NPM3D and S3DIS. Moreover, the pyramid architecture is generic
 and can be applied to other network designs: we show an example of similar impr
ovements over RandLANet.
****

Decoupling Makes Weakly Supervised Local Feature Better

Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15838-15848

Weakly supervised learning can help local feature methods to overcome the obstacle of acquiring a large-scale dataset with densely labeled correspondences. However, since weak supervision cannot distinguish the losses caused by the detection and description steps, directly conducting weakly supervised learning within a joint training describe-then-detect pipeline suffers limited performance. In this paper, we propose a decoupled training describe-then-detect pipeline tailored for weakly supervised local feature learning. Within our pipeline, the detection step is decoupled from the description step and postponed until discriminative and robust descriptors are learned. In addition, we introduce a line-to-window search strategy to explicitly use the camera pose information for better descriptor learning. Extensive experiments show that our method, namely PoSFeat (Camera Pose Supervised Feature), outperforms previous fully and weakly supervised methods and achieves state-ofthe-art performance on a wide range of downstream tasks.

****

Cross-Architecture Self-Supervised Video Representation Learning

Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, Weilin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19270-19279

In this paper, we present a new cross-architecture contrastive learning (CACL) framework for self-supervised video representation learning. CACL consists of a 3D CNN and a video transformer which are used in parallel to generate diverse positive pairs for contrastive learning. This allows the model to learn strong representations from such diverse yet meaningful pairs. Furthermore, we introduce a temporal self-supervised learning module able to predict an Edit distance explicitly between two video sequences in the temporal order. This enables the model to learn a rich temporal representation that compensates strongly to the video-level representation learned by the CACL. We evaluate our method on the tasks of video retrieval and action recognition on UCF101 and HMDB51 datasets, where our method achieves excellent performance, surpassing the state-of-the-art methods such as VideoMoCo and MoCo+BE by a large margin.

****

High-Resolution Image Harmonization via Collaborative Dual Transformations

Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, Liqing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18470-18479

Given a composite image, image harmonization aims to adjust the foreground to make it compatible with the background. High-resolution image harmonization is in high demand, but still remains unexplored. Conventional image harmonization methods learn global RGB-to-RGB transformation which could effortlessly scale to high resolution, but ignore diverse local context. Recent deep learning methods learn the dense pixel-to-pixel transformation which could generate harmonious outputs, but are highly constrained in low resolution. In this work, we propose a high-resolution image harmonization network with Collaborative Dual Transformation (CDTNet) to combine pixel-to-pixel transformation and RGB-to-RGB transformation coherently in an end-to-end network. Our CDTNet consists of a low-resolution generator for pixel-to-pixel transformation, a color mapping module for RGB-to-RGB transformation, and a refinement module to take advantage of both. Extensive experiments on high-resolution benchmark dataset and our created high-resolution real composite images demonstrate that our CDTNet strikes a good balance between efficiency and effectiveness. Our used datasets can be found in https://github.com/bcmi/CDTNet-High-Resolution-Image-Harmonization.

****

Homography Loss for Monocular 3D Object Detection

Jiaqi Gu, Bojian Wu, Lubin Fan, Jianqiang Huang, Shen Cao, Zhiyu Xiang, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1080-1089

Monocular 3D object detection is an essential task in autonomous driving. Howeve r, most current methods consider each 3D object in the scene as an independent t raining sample, while ignoring their inherent geometric relations, thus inevitab ly resulting in a lack of leveraging spatial constraints. In this paper, we prop ose a novel method that takes all the objects into consideration and explores th eir mutual relationships to help better estimate the 3D boxes. Moreover, since 2 D detection is more reliable currently, we also investigate how to use the detec ted 2D boxes as guidance to globally constrain the optimization of the correspon ding predicted 3D boxes. To this end, a differentiable loss function, termed as Homography Loss, is proposed to achieve the goal, which exploits both 2D and 3D information, aiming at balancing the positional relationships between different objects by global constraints, so as to obtain more accurately predicted 3D boxe s. Thanks to the concise design, our loss function is universal and can be plugg ed into any mature monocular 3D detector, while significantly boosting the perfo rmance over their baseline. Experiments demonstrate that our method yields the b est performance (Nov. 2021) compared with the other state-of-the-arts by a large margin on KITTI 3D datasets.

****

A Unified Model for Line Projections in Catadioptric Cameras With Rotationally S ymmetric Mirrors

Pedro Miraldo, José Pedro Iglesias; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2022, pp. 15797-15806

Lines are among the most used computer vision features, in applications such as camera calibration to object detection. Catadioptric cameras with rotationally s ymmetric mirrors are omnidirectional imaging devices, capturing up to a 360 degr ees field of view. These are used in many applications ranging from robotics to panoramic vision. Although known for some specific configurations, the modeling of line projection was never fully solved for general central and non-central ca tadioptric cameras. We start by taking some general point reflection assumptions and derive a line reflection constraint. This constraint is then used to define a line projection into the image. Next, we compare our model with previous meth ods, showing that our general approach outputs the same polynomial degrees as pr evious configuration-specific systems. We run several experiments using syntheti c and real-world data, validating our line projection model. Lastly, we show an application of our methods to an absolute camera pose problem.

****

Dynamic Sparse R-CNN

Qinghang Hong, Fengming Liu, Dong Li, Ji Liu, Lu Tian, Yi Shan; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4723-4732

Sparse R-CNN is a recent strong object detection baseline by set prediction on s parse, learnable proposal boxes and proposal features. In this work, we propose to improve Sparse R-CNN with two dynamic designs. First, Sparse R-CNN adopts a o ne-to-one label assignment scheme, where the Hungarian algorithm is applied to m atch only one positive sample for each ground truth. Such one-to-one assignment may not be optimal for the matching between the learned proposal boxes and groun d truths. To address this problem, we propose dynamic label assignment (DLA) bas ed on the optimal transport algorithm to assign increasing positive samples in t he iterative training stages of Sparse R-CNN. We constrain the matching to be gr adually looser in the sequential stages as the later stage produces the refined proposals with improved precision. Second, the learned proposal boxes and featur es remain fixed for different images in the inference process of Sparse R-CNN. M otivated by dynamic convolution, we propose dynamic proposal generation (DPG) to assemble multiple proposal experts dynamically for providing better initial pro posal boxes and features for the consecutive training stages. DPG thereby can de rive sample-dependent proposal boxes and features for inference. Experiments dem onstrate that our method, named Dynamic Sparse R-CNN, can boost the strong Spars e R-CNN baseline with different backbones for object detection. Particularly, Dy namic Sparse R-CNN reaches the state-of-the-art 47.2% AP on the COCO 2017 valida tion set, surpassing Sparse R-CNN by 2.2% AP with the same ResNet-50 backbone.

****
MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation
Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Ga rg, In So Kweon, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 16928-16937
Test-time adaptation approaches have recently emerged as a practical solution fo r handling domain shift without access to the source domain data. In this paper, we propose and explore a new multi-modal extension of test-time adaptation for 3D semantic segmentation. We find that, directly applying existing methods usual ly results in performance instability at test time, because multi-modal input is not considered jointly. To design a framework that can take full advantage of m ulti-modality, where each modality provides regularized self-supervisory signals to other modalities, we propose two complementary modules within and across the modalities. First, Intra-modal Pseudo-label Generation (Intra-PG) is introduced to obtain reliable pseudo labels within each modality by aggregating informatio n from two models that are both pre-trained on source data but updated with targ et data at different paces. Second, Intermodal Pseudo-label Refinement (Inter-PR) adaptively selects more reliable pseudo labels from different modalities based on a proposed consistency scheme. Experiments demonstrate that our regularized pseudo labels produce stable self-learning signals in numerous multi-modal test- time adaptation scenarios for 3D semantic segmentation. Visit our project websit e at https://www.nec-labs.com/ mas/MM-TTA.
****
Stable Long-Term Recurrent Video Super-Resolution
Benjamin Naoto Chiche, Arnaud Woiselle, Joana Frontera-Pons, Jean-Luc Starck; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 837-846
Recurrent models have gained popularity in deep learning (DL) based video super- resolution (VSR), due to their increased computational efficiency, temporal rece ptive field and temporal consistency compared to sliding-window based models. Ho wever, when inferring on long video sequences presenting low motion (i.e. in whi ch some parts of the scene barely move), recurrent models diverge through recurr ent processing, generating high frequency artifacts. To the best of our knowledg e, no study about VSR pointed out this instability problem, which can be critica l for some real-world applications. Video surveillance is a typical example wher e such artifacts would occur, as both the camera and the scene stay static for a long time. In this work, we expose instabilities of existing recurrent VSR netw orks on long sequences with low motion. We demonstrate it on a new long sequence dataset Quasi-Static Video Set, that we have created. Finally, we introduce a n ew framework of recurrent VSR networks that is both stable and competitive, base d on Lipschitz stability theory. We propose a new recurrent VSR network, coined Middle Recurrent Video Super-Resolution (MRVSR), based on this framework. We emp irically show its competitive performance on long sequences with low motion.
****
Dual-Generator Face Reenactment
Gee-Sern Hsu, Chun-Hung Tsai, Hung-Yi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 642-650
We propose the Dual-Generator (DG) network for large-pose face reenactment. Give n a source face and a reference face as inputs, the DG network can generate an o utput face that has the same pose and expression as of the reference face, and h as the same identity as of the source face. As most approaches do not particular ly consider large-pose reenactment, the proposed approach addresses this issue b y incorporating a 3D landmark detector into the framework and considering a loss function to capture visible local shape variation across large pose. The DG net work consists of two modules, the ID-preserving Shape Generator (IDSG) and the R eenacted Face Generator (RFG). The IDSG encodes the 3D landmarks of the referenc e face into a reference landmark code, and encodes the source face into a source face code. The reference landmark code and the source face code are concatenate d and decoded to a set of target landmarks that exhibits the pose and expression of the reference face and preserves the identity of the source face.

****
Towards Bidirectional Arbitrary Image Rescaling: Joint Optimization and Cycle Id empotence
Zhihong Pan, Baopu Li, Dongliang He, Mingde Yao, Wenhao Wu, Tianwei Lin, Xin Li, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 17389-17398
Deep learning based single image super-resolution models have been widely studie d and superb results are achieved in upscaling low-resolution images with fixed scale factor and downscaling degradation kernel. To improve real world applicabi lity of such models, there are growing interests to develop models optimized for arbitrary upscaling factors. Our proposed method is the first to treat arbitrar y rescaling, both upscaling and downscaling, as one unified process. Using joint optimization of both directions, the proposed model is able to learn upscaling and downscaling simultaneously and achieve bidirectional arbitrary image rescali ng. It improves the performance of current arbitrary upscaling models by a large margin while at the same time learns to maintain visual perception quality in d ownscaled images. The proposed model is further shown to be robust in cycle idem potence test, free of severe degradations in reconstruction accuracy when the do wnscaling-to-upscaling cycle is applied repetitively. This robustness is benefic ial for image rescaling in the wild when this cycle could be applied to one imag e for multiple times. It also performs well on tests with arbitrary large scales and asymmetric scales, even when the model is not trained with such tasks. Exte nsive experiments are conducted to demonstrate the superior performance of our m odel.
****
Self-Supervised Neural Articulated Shape and Appearance Models
Fangyin Wei, Rohan Chabra, Lingni Ma, Christoph Lassner, Michael Zollhöfer, Szym on Rusinkiewicz, Chris Sweeney, Richard Newcombe, Mira Slavcheva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 15816-15826
Learning geometry, motion, and appearance priors of object classes is important for the solution of a large variety of computer vision problems. While the major ity of approaches has focused on static objects, dynamic objects, especially wit h controllable articulation, are less explored. We propose a novel approach for learning a representation of the geometry, appearance, and motion of a class of articulated objects given only a set of color images as input. In a self-supervi sed manner, our novel representation learns shape, appearance, and articulation codes that enable independent control of these semantic dimensions. Our model is trained end-to-end without requiring any articulation annotations. Experiments show that our approach performs well for different joint types, such as revolute and prismatic joints, as well as different combinations of these joints. Compar ed to state of the art that uses direct 3D supervision and does not output appea rance, we recover more faithful geometry and appearance from 2D observations onl y. In addition, our representation enables a large variety of applications, such as few-shot reconstruction, the generation of novel articulations, and novel vi ew-synthesis. Project page: https://weify627.github.io/nasam/.
****
A Hybrid Quantum-Classical Algorithm for Robust Fitting
Anh-Dzung Doan, Michele Sasdelli, David Suter, Tat-Jun Chin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 417-427
Fitting geometric models onto outlier contaminated data is provably intractable. Many computer vision systems rely on random sampling heuristics to solve robust fitting, which do not provide optimality guarantees and error bounds. It is the refore critical to develop novel approaches that can bridge the gap between exac t solutions that are costly, and fast heuristics that offer no quality assurance s. In this paper, we propose a hybrid quantum-classical algorithm for robust fit ting. Our core contribution is a novel robust fitting formulation that solves a sequence of integer programs and terminates with a global solution or an error b ound. The combinatorial subproblems are amenable to a quantum annealer, which he

lps to tighten the bound efficiently. While our usage of quantum computing does not surmount the fundamental intractability of robust fitting, by providing error bounds our algorithm is a practical improvement over randomised heuristics. Moreover, our work represents a concrete application of quantum computing in computer vision. We present results obtained using an actual quantum computer (D-Wave Advantage) and via simulation.
****

Topology Preserving Local Road Network Estimation From Single Onboard Camera Image
Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17263-17272
Knowledge of the road network topology is crucial for autonomous planning and navigation. Yet, recovering such topology from a single image has only been explored in part. Furthermore, it needs to refer to the ground plane, where also the driving actions are taken. This paper aims at extracting the local road network topology, directly in the bird's-eye-view (BEV), all in a complex urban setting. The only input consists of a single onboard, forward looking camera image. We represent the road topology using a set of directed lane curves and their interactions, which are captured using their intersection points. To better capture topology, we introduce the concept of minimal cycles and their covers. A minimal cycle is the smallest cycle formed by the directed curve segments (between two intersections). The cover is a set of curves whose segments are involved in forming a minimal cycle. We first show that the covers suffice to uniquely represent the road topology. The covers are then used to supervise deep neural networks, along with the lane curve supervision. These learn to predict the road topology from a single input image. The results on the NuScenes and Argoverse benchmarks are significantly better than those obtained with baselines. Code: https://github.com/ybarancan/TopologicalLaneGraph.
****

Eigenlanes: Data-Driven Lane Descriptors for Structurally Diverse Lanes
Dongkwon Jin, Wonhui Park, Seong-Gyun Jeong, Heeyeon Kwon, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17163-17171
A novel algorithm to detect road lanes in the eigenlane space is proposed in this paper. First, we introduce the notion of eigenlanes, which are data-driven descriptors for structurally diverse lanes, including curved, as well as straight, lanes. To obtain eigenlanes, we perform the best rank-M approximation of a lane matrix containing all lanes in a training set. Second, we generate a set of lane candidates by clustering the training lanes in the eigenlane space. Third, using the lane candidates, we determine an optimal set of lanes by developing an anchor-based detection network, called SIIC-Net. Experimental results demonstrate that the proposed algorithm provides excellent detection performance for structurally diverse lanes. Our codes are available at https://github.com/dongkwonjin/Eigenlanes.
****

Human Instance Matting via Mutual Guidance and Multi-Instance Refinement
Yanan Sun, Chi-Keung Tang, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2647-2656
This paper introduces a new matting task called human instance matting (HIM), which requires the pertinent model to automatically predict a precise alpha matte for each human instance. Straightforward combination of closely related techniques, namely, instance segmentation, soft segmentation and human/conventional matting, will easily fail in complex cases requiring disentangling mingled colors belonging to multiple instances along hairy and thin boundary structures. To tackle these technical challenges, we propose a human instance matting framework, called InstMatt, where a novel mutual guidance strategy working in tandem with a multi-instance refinement module is used, for delineating multi-instance relationship among humans with complex and overlapping boundaries if present. A new instance matting metric called instance matting quality (IMQ) is proposed, which addr

esses the absence of a unified and fair means of evaluation emphasizing both ins tance recognition and mat-ting quality. Finally, we construct a HIM benchmark fo r evaluation, which comprises of both synthetic and natural benchmark images. In addition to thorough experimental results on HIM, preliminary results are prese nted on general instance matting beyond multiple and overlapping human instances .

****

TCTrack: Temporal Contexts for Aerial Tracking

Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, Changhong Fu; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 14798-14808

Temporal contexts among consecutive frames are far from being fully utilized in existing visual trackers. In this work, we present TCTrack, a comprehensive fram ework to fully exploit temporal contexts for aerial tracking. The temporal conte xts are incorporated at two levels: the extraction of features and the refinemen t of similarity maps. Specifically, for feature extraction, an online temporally adaptive convolution is proposed to enhance the spatial features using temporal information, which is achieved by dynamically calibrating the convolution weigh ts according to the previous frames. For similarity map refinement, we propose a n adaptive temporal transformer, which first effectively encodes temporal knowle dge in a memory-efficient way, before the temporal knowledge is decoded for accu rate adjustment of the similarity map. TCTrack is effective and efficient: evalu ation on four aerial tracking benchmarks shows its impressive performance; real- world UAV tests show its high speed of over 27 FPS on NVIDIA Jetson AGX Xavier.

****

SpaceEdit: Learning a Unified Editing Space for Open-Domain Image Color Editing

Jing Shi, Ning Xu, Haitian Zheng, Alex Smith, Jiebo Luo, Chenliang Xu; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19730-19739

Recently, large pretrained models (e.g., BERT, StyleGAN, CLIP) show great knowle dge transfer and generalization capability on various downstream tasks within th eir domains. Inspired by these efforts, in this paper we propose a unified model for open-domain image editing focusing on color and tone adjustment of open-dom ain images while keeping their original content and structure. Our model learns a unified editing space that is more semantic, intuitive, and easy to manipulate than the operation space (e.g., contrast, brightness, color curve) used in many existing photo editing softwares. Our model belongs to the image-to-image trans lation framework which consists of an image encoder and decoder, and is trained on pairs of before-and-after edited images to produce multimodal outputs. We sho w that by inverting image pairs into latent codes of the learned editing space, our model can be leveraged for various downstream editing tasks such as language -guided image editing, personalized editing, editing-style clustering, retrieval , etc. We extensively study the unique properties of the editing space in experi ments and demonstrate superior performance on the aforementioned tasks.

****

GAN-Supervised Dense Visual Alignment

William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A. Efros, Eli Shechtman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 13470-13481

We propose GAN-Supervised Learning, a framework for learning discriminative mode ls and their GAN-generated training data jointly end-to-end. We apply our framew ork to the dense visual alignment problem. Inspired by the classic Congealing me thod, our GANgealing algorithm trains a Spatial Transformer to map random sample s from a GAN trained on unaligned data to a common, jointly-learned target mode. We show results on eight datasets, all of which demonstrate our method successf ully aligns complex data and discovers dense correspondences. GANgealing signifi cantly outperforms past self-supervised correspondence algorithms and performs o n-par with (and sometimes exceeds) state-of-the-art supervised correspondence al gorithms on several datasets---without making use of any correspondence supervis ion or data augmentation and despite being trained exclusively on GAN-generated

data. For precise correspondence, we improve upon state-of-the-art supervised methods by as much as 3x. We show applications of our method for augmented reality, image editing and automated pre-processing of image datasets for downstream GAN training.
****
SwinTextSpotter: Scene Text Spotting via Better Synergy Between Text Detection and Text Recognition
Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, Lianwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4593-4603
End-to-end scene text spotting has attracted great attention in recent years due to the success of excavating the intrinsic synergy of the scene text detection and recognition. However, recent state-of-the-art methods usually incorporate detection and recognition simply by sharing the backbone, which does not directly take advantage of the feature interaction between the two tasks. In this paper, we propose a new end-to-end scene text spotting framework termed SwinTextSpotter. Using a transformer encoder with dynamic head as the detector, we unify the two tasks with a novel Recognition Conversion mechanism to explicitly guide text localization through recognition loss. The straightforward design results in a concise framework that requires neither additional rectification module nor character-level annotation for the arbitrarily-shaped text. Qualitative and quantitative experiments on multi-oriented datasets RoIC13 and ICDAR 2015, arbitrarily-shaped datasets Total-Text and CTW1500, and multi-lingual datasets ReCTS (Chinese) and VinText (Vietnamese) demonstrate SwinTextSpotter significantly outperforms existing methods. Code is available at https://github.com/mxin262/SwinTextSpotter.
****
Multi-Level Feature Learning for Contrastive Multi-View Clustering
Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, Lifang He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16051-16060
Multi-view clustering can explore common semantics from multiple views and has attracted increasing attention. However, existing works punish multiple objectives in the same feature space, where they ignore the conflict between learning consistent common semantics and reconstructing inconsistent view-private information. In this paper, we propose a new framework of multi-level feature learning for contrastive multi-view clustering to address the aforementioned issue. Our method learns different levels of features from the raw features, including low-level features, high-level features, and semantic labels/features in a fusion-free manner, so that it can effectively achieve the reconstruction objective and the consistency objectives in different feature spaces. Specifically, the reconstruction objective is conducted on the low-level features. Two consistency objectives based on contrastive learning are conducted on the high-level features and the semantic labels, respectively. They make the high-level features effectively explore the common semantics and the semantic labels achieve the multi-view clustering. As a result, the proposed framework can reduce the adverse influence of view-private information. Extensive experiments on public datasets demonstrate that our method achieves state-of-the-art clustering effectiveness.
****
RendNet: Unified 2D/3D Recognizer With Latent Space Rendering
Ruoxi Shi, Xinyang Jiang, Caihua Shan, Yansen Wang, Dongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5408-5417
Vector graphics (VG) have been ubiquitous in our daily life with vast applications in engineering, architecture, designs, etc. The VG recognition process of most existing methods is to first render the VG into raster graphics (RG) and then conduct recognition based on RG formats. However, this procedure discards the structure of geometries and loses the high resolution of VG. Recently, another category of algorithms is proposed to recognize directly from the original VG format. But it is affected by the topological errors that can be filtered out by RG r

endering. Instead of looking at one format, it is a good solution to utilize the formats of VG and RG together to avoid these shortcomings. Besides, we argue that the VG-to-RG rendering process is essential to effectively combine VG and RG information. By specifying the rules on how to transfer VG primitives to RG pixels, the rendering process depicts the interaction and correlation between VG and RG. As a result, we propose RenderNet, a unified architecture for recognition on both 2D and 3D scenarios, which considers both VG/RG representations and exploits their interaction by incorporating the VG-to-RG rasterization process. Experiments show that RenderNet can achieve state-of-the-art performance on 2D and 3D object recognition tasks on various VG datasets.
****

iPLAN: Interactive and Procedural Layout Planning
Feixiang He, Yanlong Huang, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7793-7802
Layout design is ubiquitous in many applications, e.g. architecture/urban planning, etc, which involves a lengthy iterative design process. Recently, deep learning has been leveraged to automatically generate layouts via image generation, showing a huge potential to free designers from laborious routines. While automatic generation can greatly boost productivity, designer input is undoubtedly crucial. An ideal AI-aided design tool should automate repetitive routines, and meanwhile accept human guidance and provide smart/proactive suggestions. However, the capability of involving humans into the loop has been largely ignored in existing methods which are mostly end-to-end approaches. To this end, we propose a new human-in-the-loop generative model, iPLAN, which is capable of automatically generating layouts, but also interacting with designers throughout the whole procedure, enabling humans and AI to co-evolve a sketchy idea gradually into the final design. iPLAN is evaluated on diverse datasets and compared with existing methods. The results show that iPLAN has high fidelity in producing similar layouts to those from human designers, great flexibility in accepting designer inputs and providing design suggestions accordingly, and strong generalizability when facing unseen design tasks and limited training data.
****

Video Frame Interpolation With Transformer
Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3532-3542
Video frame interpolation (VFI), which aims to synthesize intermediate frames of a video, has made remarkable progress with development of deep convolutional networks over past years. Existing methods built upon convolutional networks generally face challenges of handling large motion due to the locality of convolution operations. To overcome this limitation, we introduce a novel framework, which takes advantage of Transformer to model long-range pixel correlation among video frames. Further, our network is equipped with a novel cross-scale window-based attention mechanism, where cross-scale windows interact with each other. This design effectively enlarges the receptive field and aggregates multi-scale information. Extensive quantitative and qualitative experiments demonstrate that our method achieves new state-of-the-art results on various benchmarks.
****

GIFS: Neural Implicit Function for General Shape Representation
Jianglong Ye, Yuntao Chen, Naiyan Wang, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12829-12839
Recent development of neural implicit function has shown tremendous success on high-quality 3D shape reconstruction. However, most works divide the space into inside and outside of the shape, which limits their representing power to single-layer and watertight shapes. This limitation leads to tedious data processing (converting non-watertight raw data to watertight) as well as the incapability of representing general object shapes in the real world. In this work, we propose a novel method to represent general shapes including non-watertight shapes and shapes with multi-layer surfaces. We introduce General Implicit Function for 3D Sh

ape (GIFS), which models the relationships between every two points instead of t
he relationships between points and surfaces. Instead of dividing 3D space into
predefined inside-outside regions, GIFS encodes whether two points are separated
 by any surface. Experiments on ShapeNet show that GIFS outperforms previous sta
te-of-the-art methods in terms of reconstruction quality, rendering efficiency,
and visual fidelity.
****

Deblur-NeRF: Neural Radiance Fields From Blurry Images
Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, Pedro V. Sander; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2022, pp. 12861-12870
Neural Radiance Field (NeRF) has gained considerable attention recently for 3D s
cene reconstruction and novel view synthesis due to its remarkable synthesis qua
lity. However, image blurriness caused by defocus or motion, which often occurs
when capturing scenes in the wild, significantly degrades its reconstruction qua
lity. To address this problem, We propose Deblur-NeRF, the first method that can
 recover a sharp NeRF from blurry input. We adopt an analysis-by-synthesis appro
ach that reconstructs blurry views by simulating the blurring process, thus maki
ng NeRF robust to blurry inputs. The core of this simulation is a novel Deformab
le Sparse Kernel (DSK) module that models spatially-varying blur kernels by defo
rming a canonical sparse kernel at each spatial location. The ray origin of each
 kernel point is jointly optimized, inspired by the physical blurring process. T
his module is parameterized as an MLP that has the ability to be generalized to
various blur types. Jointly optimizing the NeRF and the DSK module allows us to
restore a sharp NeRF. We demonstrate that our method can be used on both camera
motion blur and defocus blur: the two most common types of blur in real scenes.
Evaluation results on both synthetic and real-world data show that our method ou
tperforms several baselines. The synthetic and real datasets along with the sour
ce code can be find in https://limacv.github.io/deblurnerf/
****

Egocentric Prediction of Action Target in 3D
Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, Chen
 Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2022, pp. 21003-21012
We are interested in anticipating as early as possible the target location of a
person's object manipulation action in a 3D workspace from egocentric vision. It
 is important in fields like human-robot collaboration, but has not yet received
 enough attention from vision and learning communities. To stimulate more resear
ch on this challenging egocentric vision task, we propose a large multimodality
dataset of more than 1 million frames of RGB-D and IMU streams, and provide eval
uation metrics based on our high-quality 2D and 3D labels from semi-automatic an
notation. Meanwhile, we design baseline methods using recurrent neural networks
and conduct various ablation studies to validate their effectiveness. Our result
s demonstrate that this new task is worthy of further study by researchers in ro
botics, vision, and learning communities.
****

TemporalUV: Capturing Loose Clothing With Temporally Coherent UV Coordinates
You Xie, Huiqi Mao, Angela Yao, Nils Thuerey; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3450-3459
We propose a novel approach to generate temporally coherent UV coordinates for l
oose clothing. Our method is not constrained by human body outlines and can capt
ure loose garments and hair. We implemented a differentiable pipeline to learn U
V mapping between a sequence of RGB inputs and textures via UV coordinates. Inst
ead of treating the UV coordinates of each frame separately, our data generation
 approach connects all UV coordinates via feature matching for temporal stabilit
y. Subsequently, a generative model is trained to balance the spatial quality an
d temporal stability. It is driven by supervised and unsupervised losses in both
 UV and image spaces. Our experiments show that the trained models output high-q
uality UV coordinates and generalize to new poses. Once a sequence of UV coordin
ates has been inferred by our model, it can be used to flexibly synthesize new l

ooks and modified visual styles. Compared to existing methods, our approach redu ces the computational workload to animate new outfits by several orders of magni tude.
****

Whose Track Is It Anyway? Improving Robustness to Tracking Errors With Affinity-Based Trajectory Prediction
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 657 3-6582
Multi-agent trajectory prediction is critical for planning and decision-making i n human-interactive autonomous systems, such as self-driving cars. However, most prediction models are developed separately from their upstream perception (dete ction and tracking) modules, assuming ground truth past trajectories as inputs. As a result, their performance degrades significantly when using real-world nois y tracking results as inputs. This is typically caused by the propagation of err ors from tracking to prediction, such as noisy tracks, fragments, and identity s witches. To alleviate this propagation of errors, we propose a new prediction pa radigm that uses detections and their affinity matrices across frames as inputs, removing the need for error-prone data association during tracking. Since affin ity matrices contain "soft" information about the similarity and identity of det ections across frames, making predictions directly from affinity matrices retain s strictly more information than making predictions from the tracklets generated by data association. Experiments on large-scale, real-world autonomous driving datasets show that our affinity-based prediction scheme reduces overall predicti on errors by up to 57.9%, in comparison to standard prediction pipelines that us e tracklets as inputs, with even more significant error reduction (up to 88.6%) if restricting the evaluation to challenging scenarios with tracking errors. Our project website is at https://www.xinshuoweng.com/projects/Affinipred
****

DoubleField: Bridging the Neural Surface and Radiance Fields for High-Fidelity H uman Reconstruction and Rendering
Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, Yebin L iu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 15872-15882
We introduce DoubleField, a novel framework combining the merits of both surface field and radiance field for high-fidelity human reconstruction and rendering. Within DoubleField, the surface field and radiance field are associated together by a shared feature embedding and a surface-guided sampling strategy. Moreover, a view-to-view transformer is introduced to fuse multi-view features and learn view-dependent features directly from high-resolution inputs. With the modeling power of DoubleField and the view-to-view transformer, our method significantly improves the reconstruction quality of both geometry and appearance, while suppo rting direct inference, scene-specific high-resolution finetuning, and fast rend ering. The efficacy of DoubleField is validated by the quantitative evaluations on several datasets and the qualitative results in a real-world sparse multi-vie w system, showing its superior capability for high-quality human model reconstru ction and photo-realistic free-viewpoint human rendering. Data and source code w ill be made public for the research purpose.
****

Towards Real-World Navigation With Deep Differentiable Planners
Shu Ishida, João F. Henriques; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 17327-17336
We train embodied neural networks to plan and navigate unseen complex 3D environ ments, emphasising real-world deployment. Rather than requiring prior knowledge of the agent or environment, the planner learns to model the state transitions a nd rewards. To avoid the potentially hazardous trial-and-error of reinforcement learning, we focus on differentiable planners such as Value Iteration Networks (VIN), which are trained offline from safe expert demonstrations. Although they w ork well in small simulations, we address two major limitations that hinder thei r deployment. First, we observed that current differentiable planners struggle t

o plan long-term in environments with a high branching complexity. While they sh
ould ideally learn to assign low rewards to obstacles to avoid collisions, these
 penalties are not strong enough to guarantee collision-free operation. We thus
impose a structural constraint on the value iteration, which explicitly learns t
o model impossible actions and noisy motion. Secondly, we extend the model to pl
an exploration with a limited perspective camera under translation and fine rota
tions, which is crucial for real robot deployment. Our proposals significantly i
mprove semantic navigation and exploration on several 2D and 3D environments, su
cceeding in settings that are otherwise challenging for differentiable planners.
 As far as we know, we are the first to successfully apply them to the difficult
 Active Vision Dataset, consisting of real images captured from a robot.
****

An Iterative Quantum Approach for Transformation Estimation From Point Sets
Natacha Kuete Meli, Florian Mannel, Jan Lellmann; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 529-537
We propose an iterative method for estimating rigid transformations from point s
ets using adiabatic quantum computation. Compared to existing quantum approaches
, our method relies on an adaptive scheme to solve the problem to high precision
, and does not suffer from inconsistent rotation matrices. Experimentally, our m
ethod performs robustly on several 2D and 3D datasets even with high outlier rat
io.
****

Video K-Net: A Simple, Strong, and Unified Baseline for Video Segmentation
Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai To
ng, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2022, pp. 18847-18857
This paper presents Video K-Net, a simple, strong, and unified framework for ful
ly end-to-end video panoptic segmentation. The method is built upon K-Net, a met
hod that unifies image segmentation via a group of learnable kernels. We observe
 that these learnable kernels from K-Net, which encode object appearances and co
ntexts, can naturally associate identical instances across video frames. Motivat
ed by this observation, Video K-Net learns to simultaneously segment and track "
things" and "stuff" in a video with simple kernel-based appearance modeling and
cross-temporal kernel interaction. Despite the simplicity, it achieves state-of-
the-art video panoptic segmentation results on Citscapes-VPS and KITTI-STEP with
out bells and whistles. In particular on KITTI-STEP, the simple method can boost
 almost 12% relative improvements over previous methods. We also validate its ge
neralization on video semantic segmentation, where we boost various baselines by
 2% on the VSPW dataset. Moreover, we extend K-Net into clip-level video framewo
rk for video instance segmentation where we obtain 40.5% for ResNet50 backbone a
nd 51.5% mAP for Swin-base on YouTube-2019 validation set. We hope this simple y
et effective method can serve as a new flexible baseline in video segmentation.
Both code and models are released at \href https://github.com/lxtGH/Video-K-Net.
****

UnweaveNet: Unweaving Activity Stories
Will Price, Carl Vondrick, Dima Damen; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13770-13779
Our lives can be seen as a complex weaving of activities; we switch from one act
ivity to another, to maximise our achievements or in reaction to demands placed
upon us. Observing a video of unscripted daily activities, we parse the video in
to its constituent activity threads through a process we call unweaving. To acco
mplish this, we introduce a video representation explicitly capturing activity t
hreads called a thread bank, along with a neural controller capable of detecting
 goal changes and continuations of past activities, together forming UnweaveNet.
 We train and evaluate UnweaveNet on sequences from the unscripted egocentric da
taset EPIC-KITCHENS. We propose and showcase the efficacy of pretraining Unweave
Net in a self-supervised manner.
****

Balanced MSE for Imbalanced Visual Regression
Jiawei Ren, Mingyuan Zhang, Cunjun Yu, Ziwei Liu; Proceedings of the IEEE/CVF Co

nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7926-7935
Data imbalance exists ubiquitously in real-world visual regressions, e.g., age estimation and pose estimation, hurting the model's generalizability and fairness. Thus, imbalanced regression gains increasing research attention recently. Compared to imbalanced classification, imbalanced regression focuses on continuous labels, which can be boundless and high-dimensional and hence more challenging. In this work, we identify that the widely used Mean Square Error (MSE) loss function can be ineffective in imbalanced regression. We revisit MSE from a statistical view and propose a novel loss function, Balanced MSE, to accommodate the imbalanced training label distribution. We further design multiple implementations of Balanced MSE to tackle different real-world scenarios, particularly including the one that requires no prior knowledge about the training label distribution. Moreover, to the best of our knowledge, Balanced MSE is the first general solution to high-dimensional imbalanced regression. Extensive experiments on both synthetic and three real-world benchmarks demonstrate the effectiveness of Balanced MSE.

****

Local Learning Matters: Rethinking Data Heterogeneity in Federated Learning
Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8397-8406
Federated learning (FL) is a promising strategy for performing privacy-preserving, distributed learning with a network of clients (i.e., edge devices). However, the data distribution among clients is often non-IID in nature, making efficient optimization difficult. To alleviate this issue, many FL algorithms focus on mitigating the effects of data heterogeneity across clients by introducing a variety of proximal terms, some incurring considerable compute and/or memory overheads, to restrain local updates with respect to the global model. Instead, we consider rethinking solutions to data heterogeneity in FL with a focus on local learning generality rather than proximal restriction. To this end, we first present a systematic study informed by second-order indicators to better understand algorithm effectiveness in FL. Interestingly, we find that standard regularization methods are surprisingly strong performers in mitigating data heterogeneity effects. Based on our findings, we further propose a simple and effective method, FedAlign, to overcome data heterogeneity and the pitfalls of previous methods. FedAlign achieves competitive accuracy with state-of-the-art FL methods across a variety of settings while minimizing computation and memory overhead. Code is available at https://github.com/mmendiet/FedAlign.

****

PhysFormer: Facial Video-Based Physiological Measurement With Temporal Difference Transformer
Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H.S. Torr, Guoying Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4186-4196
Remote photoplethysmography (rPPG), which aims at measuring heart activities and physiological signals from facial video without any contact, has great potential in many applications. Recent deep learning approaches focus on mining subtle rPPG clues using convolutional neural networks with limited spatio-temporal receptive fields, which neglect the long-range spatio-temporal perception and interaction for rPPG modeling. In this paper, we propose the PhysFormer, an end-to-end video transformer based architecture, to adaptively aggregate both local and global spatio-temporal features for rPPG representation enhancement. As key modules in PhysFormer, the temporal difference transformers first enhance the quasi-periodic rPPG features with temporal difference guided global attention, and then refine the local spatio-temporal representation against interference. Furthermore, we also propose the label distribution learning and a curriculum learning inspired dynamic constraint in frequency domain, which provide elaborate supervisions for PhysFormer and alleviate overfitting. Comprehensive experiments are performed on four benchmark datasets to show our superior performance on both intra- and cross-dataset testings. One highlight is that, unlike most transformer networ

ks needed pretraining from large-scale datasets, the proposed PhysFormer can be easily trained from scratch on rPPG datasets, which makes it promising as a novel transformer baseline for the rPPG community. The codes will be released soon.
****

Dimension Embeddings for Monocular 3D Object Detection
Yunpeng Zhang, Wenzhao Zheng, Zheng Zhu, Guan Huang, Dalong Du, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1589-1598
Most existing deep learning-based approaches for monocular 3D object detection directly regress the dimensions of objects and overlook their importance in solving the ill-posed problem. In this paper, we propose a general method to learn appropriate embeddings for dimension estimation in monocular 3D object detection. Specifically, we consider two intuitive clues in learning the dimension-aware embeddings with deep neural networks. First, we constrain the pair-wise distance on the embedding space to reflect the similarity of corresponding dimensions so that the model can take advantage of inter-object information to learn more discriminative embeddings for dimension estimation. Second, we propose to learn representative shape templates on the dimension-aware embedding space. Through the attention mechanism, each object can interact with the learnable templates and obtain the attentive dimensions as the initial estimation, which is further refined by the combined features from both the object and the attentive templates. Experimental results on the well-established KITTI dataset demonstrate the proposed method of dimension embeddings can bring consistent improvements with negligible computation cost overhead. We achieve new state-of-the-art performance on the KITTI 3D object detection benchmark.
****

Look Closer To Supervise Better: One-Shot Font Generation via Component-Based Discriminator
Yuxin Kong, Canjie Luo, Weihong Ma, Qiyuan Zhu, Shenggao Zhu, Nicholas Yuan, Lianwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13482-13491
Automatic font generation remains a challenging research issue due to the large amounts of characters with complicated structures. Typically, only a few samples can serve as the style/content reference (termed few-shot learning), which further increases the difficulty to preserve local style patterns or detailed glyph structures. We investigate the drawbacks of previous studies and find that a coarse-grained discriminator is insufficient for supervising a font generator. To this end, we propose a novel Component-Aware Module (CAM), which supervises the generator to decouple content and style at a more fine-grained level, i.e., the component level. Different from previous studies struggling to increase the complexity of generators, we aim to perform more effective supervision for a relatively simple generator to achieve its full potential, which is a brand new perspective for font generation. The whole framework achieves remarkable results by coupling component-level supervision with adversarial learning, hence we call it Component-Guided GAN, shortly CG-GAN. Extensive experiments show that our approach outperforms state-of-the-art one-shot font generation methods. Furthermore, it can be applied to handwritten word synthesis and scene text image editing, suggesting the generalization of our approach.
****

NeRFReN: Neural Radiance Fields With Reflections
Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, Song-Hai Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18409-18418
Neural Radiance Fields (NeRF) has achieved unprecedented view synthesis quality using coordinate-based neural scene representations. However, NeRF's view dependency can only handle simple reflections like highlights but cannot deal with complex reflections such as those from glass and mirrors. In these scenarios, NeRF models the virtual image as real geometries which leads to inaccurate depth estimation, and produces blurry renderings when the multi-view consistency is violated as the reflected objects may only be seen under some of the viewpoints. To ov

ercome these issues, we introduce NeRFReN, which is built upon NeRF to model scenes with reflections. Specifically, we propose to split a scene into transmitted and reflected components, and model the two components with separate neural radiance fields. Considering that this decomposition is highly under-constrained, we exploit geometric priors and apply carefully-designed training strategies to achieve reasonable decomposition results. Experiments on various self-captured scenes show that our method achieves high-quality novel view synthesis and physically sound depth estimation results while enabling scene editing applications.
****

Blind Image Super-Resolution With Elaborate Degradation Modeling on Noise and Kernel

Zongsheng Yue, Qian Zhao, Jianwen Xie, Lei Zhang, Deyu Meng, Kwan-Yee K. Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2128-2138

While researches on model-based blind single image super-resolution (SISR) have achieved tremendous successes recently, most of them do not consider the image degradation sufficiently. Firstly, they always assume image noise obeys an independent and identically distributed (i.i.d.) Gaussian or Laplacian distribution, which largely underestimates the complexity of real noise. Secondly, previous commonly-used kernel priors (e.g., normalization, sparsity) are not effective enough to guarantee a rational kernel solution, and thus degenerates the performance of subsequent SISR task. To address the above issues, this paper proposes a model-based blind SISR method under the probabilistic framework, which elaborately models image degradation from the perspectives of noise and blur kernel. Specifically, instead of the traditional i.i.d. noise assumption, a patch-based non-i.i.d. noise model is proposed to tackle the complicated real noise, expecting to increase the degrees of freedom of the model for noise representation. As for the blur kernel, we novelly con- struct a concise yet effective kernel generator, and plug it into the proposed blind SISR method as an explicit kernel prior (EKP). To solve the proposed model, a theoretically grounded Monte Carlo EM algorithm is specifically designed. Comprehensive experiments demonstrate the superiority of our method over current state-of-the-arts on synthetic and real datasets. The source code is available at https://github.com/zsyOAOA/BSRDM.
****

Finding Good Configurations of Planar Primitives in Unorganized Point Clouds

Mulin Yu, Florent Lafarge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6367-6376

We present an algorithm for detecting planar primitives from unorganized 3D point clouds. Departing from an initial configuration, the algorithm refines both the continuous plane parameters and the discrete assignment of input points to them by seeking high fidelity, high simplicity and high completeness. Our key contribution relies upon the design of an exploration mechanism guided by a multi-objective energy function. The transitions within the large solution space are handled by five geometric operators that create, remove and modify primitives. We demonstrate the potential of our method on a variety of scenes, from organic shapes to man-made objects, and sensors, from multiview stereo to laser. We show its efficacy with respect to existing primitive fitting approaches and illustrate its applicative interest in compact mesh reconstruction, when combined with a plane assembly method.
****

PhyIR: Physics-Based Inverse Rendering for Panoramic Indoor Images

Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, Jiaqi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12713-12723

Inverse rendering of complex material such as glossy, metal and mirror material is a long-standing ill-posed problem in this area, which has not been well solved. Previous approaches cannot tackle them well due to simplified BRDF and unsuitable illumination representations. In this paper, we present PhyIR, a neural inverse rendering method with a more completed SVBRDF representation and a physics-based in-network rendering layer, which can handle complex material and incorpor

ate physical constraints by re-rendering realistic and detailed specular reflect ance. Our framework estimates geometry, material and Spatially-Coherent (SC) ill umination from a single indoor panorama. Due to the lack of panoramic datasets w ith completed SVBRDF and full-spherical light probes, we introduce an artist-des igned dataset named FutureHouse with high-quality geometry, SVBRDF and per-pixel Spatially-Varying (SV) lighting. To ensure the coherence of SV lighting, a nove l SC loss is proposed. Extensive experiments on both synthetic and real-world da ta show that the proposed method outperforms the state-of-the-arts quantitativel y and qualitatively, and is able to produce photorealistic results for a number of applications such as dynamic virtual object insertion.
****

SCS-Co: Self-Consistent Style Contrastive Learning for Image Harmonization
Yucheng Hang, Bin Xia, Wenming Yang, Qingmin Liao; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19710-197 19
Image harmonization aims to achieve visual consistency in composite images by ad apting a foreground to make it compatible with a background. However, existing m ethods always only use the real image as the positive sample to guide the traini ng, and at most introduce the corresponding composite image as a single negative sample for an auxiliary constraint, which leads to limited distortion knowledge , and further causes a too large solution space, making the generated harmonized image distorted. Besides, none of them jointly constrain from the foreground se lf-style and foreground-background style consistency, which exacerbates this pro blem. Moreover, recent region-aware adaptive instance normalization achieves gre at success but only considers the global background feature distribution, making the aligned foreground feature distribution biased. To address these issues, we propose a self-consistent style contrastive learning scheme (SCS-Co). By dynami cally generating multiple negative samples, our SCS-Co can learn more distortion knowledge and well regularize the generated harmonized image in the style repre sentation space from two aspects of the foreground self-style and foreground-bac kground style consistency, leading to a more photorealistic visual result. In ad dition, we propose a background-attentional adaptive instance normalization (BAI N) to achieve an attention-weighted background feature distribution according to the foreground-background feature similarity. Experiments demonstrate the super iority of our method over other state-of-the-art methods in both quantitative co mparison and visual analysis.
****

Beyond Fixation: Dynamic Window Visual Transformer
Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, Xiao jun Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11987-11997
Recently, a surge of interest in visual transformers is to reduce the computatio nal cost by limiting the calculation of self-attention to a local window. Most c urrent work uses a fixed single-scale window for modeling by default, ignoring t he impact of window size on model performance. However, this may limit the model ing potential of these window-based models for multi-scale information. In this paper, we propose a novel method, named Dynamic Window Vision Transformer (DW-Vi T). To the best of our knowledge, we are the first to use dynamic multi-scale wi ndows to explore the upper limit of the effect of window settings on model perfo rmance. In DW-ViT, multi-scale information is obtained by assigning windows of d ifferent sizes to different head groups of window multi-head self-attention. The n, the information is dynamically fused by assigning different weights to the mu lti-scale window branches. We conducted a detailed performance evaluation on thr ee datasets, ImageNet-1K, ADE20K, and COCO. Compared with related state-of-the-a rt (SoTA) methods, DW-ViT obtains the best performance. Specifically, compared w ith the current SoTA Swin Transformers [??], DW-ViT has achieved consistent and substantial improvements on all three datasets with similar parameters and compu tational costs. In addition, DW-ViT exhibits good scalability and can be easily inserted into any window-based visual transformers.
****

Progressive End-to-End Object Detection in Crowded Scenes
Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 857-866
In this paper, we propose a new query-based detection framework for crowd detection. Previous query-based detectors suffer from two drawbacks: first, multiple predictions will be inferred for a single object, typically in crowded scenes; second, the performance saturates as the depth of the decoding stage increases. Benefiting from the nature of the one-to-one label assignment rule, we propose a progressive predicting method to address the above issues. Specifically, we first select accepted queries prone to generate true positive predictions, then refine the rest noisy queries according to the previously accepted predictions. Experiments show that our method can significantly boost the performance of query-based detectors in crowded scenes. Equipped with our approach, Sparse RCNN achieves 92.0% \text AP , 41.4% \text MR ^ -2 and 83.2% \text JI on the challenging CrowdHuman [??] dataset, outperforming the box-based method MIP [??] that specifies in handling crowded scenarios. Moreover, the proposed method, robust to crowdedness, can still obtain consistent improvements on moderately and slightly crowded datasets like CityPersons [??] and COCO [??].
****

FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification
Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, Jungong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7349-7358
For Visible-Infrared person Re-IDentification (VI-ReID), existing modality-specific information compensation based models try to generate the images of missing modality from existing ones for reducing cross-modality discrepancy. However, because of the large modality discrepancy between visible and infrared images, the generated images usually have low qualities and introduce much more interfering information (e.g., color inconsistency). This greatly degrades the subsequent VI-ReID performance. Alternatively, we present a novel Feature-level Modality Compensation Network (FMCNet) for VIReID in this paper, which aims to compensate the missing modality-specific information in the feature level rather than in the image level, i.e., directly generating those missing modality-specific features of one modality from existing modality-shared features of the other modality. This will enable our model to mainly generate some discriminative person related modality-specific features and discard those non-discriminative ones for benefiting VI-ReID. For that, a single-modality feature decomposition module is first designed to decompose single-modality features into modality-specific ones and modality-shared ones. Then, a feature-level modality compensation module is present to generate those missing modality-specific features from existing modality-shared ones. Finally, a shared-specific feature fusion module is proposed to combine the existing and generated features for VI-ReID. The effectiveness of our proposed model is verified on two benchmark datasets.
****

Improving GAN Equilibrium by Raising Spatial Awareness
Jianyuan Wang, Ceyuan Yang, Yinghao Xu, Yujun Shen, Hongdong Li, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11285-11293
The success of Generative Adversarial Networks (GANs) is largely built upon the adversarial training between a generator (G) and a discriminator (D). They are expected to reach a certain equilibrium where D cannot distinguish the generated images from the real ones. However, such an equilibrium is rarely achieved in practical GAN training, instead, D almost always surpasses G. We attribute one of its sources to the information asymmetry between D and G. We observe that D learns its own visual attention when determining whether an image is real or fake, but G has no explicit clue on which regions to focus on for a particular synthesis. To alleviate the issue of D dominating the competition in GANs, we aim to raise the spatial awareness of G. Randomly sampled multi-level heatmaps are encoded

into the intermediate layers of G as an inductive bias. Thus G can purposefully improve the synthesis of certain image regions. We further propose to align the spatial awareness of G with the attention map induced from D. Through this way we effectively lessen the information gap between D and G. Extensive results show that our method pushes the two-player game in GANs closer to the equilibrium, leading to a better synthesis performance. As a byproduct, the introduced spatial awareness facilitates interactive editing over the output synthesis.

****

Neural Convolutional Surfaces

Luca Morreale, Noam Aigerman, Paul Guerrero, Vladimir G. Kim, Niloy J. Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19333-19342

This work is concerned with representation of shapes while disentangling fine, local and possibly repeating geometry, from global, coarse structures. Achieving such disentanglement leads to two unrelated advantages: i) a significant compression in the number of parameters required to represent a given geometry; ii) the ability to manipulate either global geometry, or local details, without harming the other. At the core of our approach lies a novel pipeline and neural architecture, which are optimized to represent one specific atlas, representing one 3D surface. Our pipeline and architecture are designed so that disentanglement of global geometry from local details is accomplished through optimization, in a completely unsupervised manner. We show that this approach achieves better neural shape compression than the state of the art, as well as enabling manipulation and transfer of shape details.

****

HyperSegNAS: Bridging One-Shot Neural Architecture Search With 3D Medical Image Segmentation Using HyperNet

Cheng Peng, Andriy Myronenko, Ali Hatamizadeh, Vishwesh Nath, Md Mahfuzur Rahman Siddiquee, Yufan He, Daguang Xu, Rama Chellappa, Dong Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20741-20751

Semantic segmentation of 3D medical images is a challenging task due to the high variability of the shape and pattern of objects (such as organs or tumors). Given the recent success of deep learning in medical image segmentation, Neural Architecture Search (NAS) has been introduced to find high-performance 3D segmentation network architectures. However, because of the massive computational requirements of 3D data and the discrete optimization nature of architecture search, previous NAS methods require a long search time or necessary continuous relaxation, and commonly lead to sub-optimal network architectures. While one-shot NAS can potentially address these disadvantages, its application in the segmentation domain has not been well studied in the expansive multi-scale multi-path search space. To enable one-shot NAS for medical image segmentation, our method, named HyperSegNAS, introduces a HyperNet to assist super-net training by incorporating architecture topology information. Such a HyperNet can be removed once the super-net is trained and introduces no overhead during architecture search. We show that HyperSegNAS yields better performing and more intuitive architectures compared to the previous state-of-the-art (SOTA) segmentation networks; furthermore, it can quickly and accurately find good architecture candidates under different computing constraints. Our method is evaluated on public datasets from the Medical Segmentation Decathlon (MSD) challenge, and achieves SOTA performances.

****

A Comprehensive Study of Image Classification Model Sensitivity to Foregrounds, Backgrounds, and Visual Attributes

Mazda Moayeri, Phillip Pope, Yogesh Balaji, Soheil Feizi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19087-19097

While datasets with single-label supervision have propelled rapid advances in image classification, additional annotations are necessary in order to quantitatively assess how models make predictions. To this end, for a subset of ImageNet samples, we collect segmentation masks for the entire object and 18 informative at

tributes. We call this dataset RIVAL10 (RIch Visual Attributes with Localization), consisting of roughly 26k instances over 10 classes. Using RIVAL10, we evalua te the sensitivity of a broad set of models to noise corruptions in foregrounds, backgrounds and attributes. In our analysis, we consider diverse state-of-the-a rt architectures (ResNets, Transformers) and training procedures (CLIP, SimCLR, DeiT, Adversarial Training). We find that, somewhat surprisingly, in ResNets, ad versarial training makes models more sensitive to the background compared to for eground than standard training. Similarly, contrastively-trained models also hav e lower relative foreground sensitivity in both transformers and ResNets. Lastly , we observe intriguing adaptive abilities of transformers to increase relative foreground sensitivity as corruption level increases. Using saliency methods, we automatically discover spurious features that drive the background sensitivity of models and assess alignment of saliency maps with foregrounds. Finally, we qu antitatively study the attribution problem for neural features by comparing feat ure saliency with ground-truth localization of semantic attributes.

****

ConDor: Self-Supervised Canonicalization of 3D Pose for Partial Shapes

Rahul Sajnani, Adrien Poulenard, Jivitesh Jain, Radhika Dua, Leonidas J. Guibas, Srinath Sridhar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16969-16979

Progress in 3D object understanding has relied on manually "canonicalized" shape datasets that contain instances with consistent position and orientation (3D po se). This has made it hard to generalize these methods to in-the-wild shapes, e. g., from internet model collections or depth sensors. ConDor is a self-supervise d method that learns to Canonicalize the 3D orientation and position for full an d partial 3D point clouds. We build on top of Tensor Field Networks (TFNs), a cl ass of permutation- and rotation-equivariant, and translation-invariant 3D netwo rks. During inference, our method takes an unseen full or partial 3D point cloud at an arbitrary pose and outputs an equivariant canonical pose. During training , this network uses self-supervision losses to learn the canonical pose from an un-canonicalized collection of full and partial 3D point clouds. ConDor can also learn to consistently co-segment object parts without any supervision. Extensiv e quantitative results on four new metrics show that our approach outperforms ex isting methods while enabling new applications such as operation on depth images and annotation transfer.

****

Source-Free Domain Adaptation via Distribution Estimation

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 7212-7222

Domain Adaptation aims to transfer the knowledge learned from a labeled source d omain to an unlabeled target domain whose data distributions are different. Howe ver, the training data in source domain required by most of the existing methods is usually unavailable in real-world applications due to privacy preserving pol icies. Recently, Source-Free Domain Adaptation (SFDA) has drawn much attention, which tries to tackle domain adaptation problem without using source data. In th is work, we propose a novel framework called SFDA-DE to address SFDA task via so urce Distribution Estimation. Firstly, we produce robust pseudo-labels for targe t data with spherical k-means clustering, whose initial class centers are the we ight vectors (anchors) learned by the classifier of pretrained model. Furthermor e, we propose to estimate the class-conditioned feature distribution of source d omain by exploiting target data and corresponding anchors. Finally, we sample su rrogate features from the estimated distribution, which are then utilized to ali gn two domains by minimizing a contrastive adaptation loss function. Extensive e xperiments show that the proposed method achieves state-of-the-art performance o n multiple DA benchmarks, and even outperforms traditional DA methods which requ ire plenty of source data.

****

Robust Combination of Distributed Gradients Under Adversarial Perturbations

Kwang In Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt

ern Recognition (CVPR), 2022, pp. 254-263
We consider distributed (gradient descent-based) learning scenarios where the se rver combines the gradients of learning objectives gathered from local clients. As individual data collection and learning environments can vary, some clients c ould transfer erroneous gradients e.g., due to adversarial data or gradient pert urbations. Further, for data privacy and security, the identities of such affect ed clients are often unknown to the server. In such cases, naively aggregating t he resulting gradients can mislead the learning process. We propose a new server -side learning algorithm that robustly combines gradients. Our algorithm embeds the local gradients into the manifold of normalized gradients and refines their combinations via simulating a diffusion process therein. The resulting algorithm is instantiated as a computationally simple and efficient weighted gradient ave raging algorithm. In the experiments with five classification and three regressi on benchmark datasets, our algorithm demonstrated significant performance improv ements over existing robust gradient combination algorithms as well as the basel ine uniform gradient averaging algorithm.
****

Exploring Endogenous Shift for Cross-Domain Detection: A Large-Scale Benchmark a nd Perturbation Suppression Network
Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Bowei Jin, Hongping Zhi, Xianglong Liu, Aishan Liu; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2022, pp. 21189-21199
Existing cross-domain detection methods mostly study the domain shifts where dif ferences between domains are often caused by external environment and perceivabl e for humans. However, in real-world scenarios (e.g., MRI medical diagnosis, X-r ay security inspection), there still exists another type of shift, named endogen ous shift, where the differences between domains are mainly caused by the intrin sic factors (e.g., imaging mechanisms, hardware components, etc.), and usually i nconspicuous. This shift can also severely harm the cross-domain detection perfo rmance but has been rarely studied. To support this study, we contribute the fir st Endogenous Domain Shift (EDS) benchmark, X-ray security inspection, where the endogenous shifts among the domains are mainly caused by different X-ray machin e types with different hardware parameters, wear degrees, etc. EDS consists of 1 4,219 images including 31,654 common instances from three domains (X-ray machine s), with bounding-box annotations from 10 categories. To handle the endogenous s hift, we further introduce the Perturbation Suppression Network (PSN), motivated by the fact that this shift is mainly caused by two types of perturbations: cat egory-dependent and category-independent ones. PSN respectively exploits local p rototype alignment and global adversarial learning mechanism to suppress these t wo types of perturbations. The comprehensive evaluation results show that PSN ou tperforms SOTA methods, serving a new perspective to the cross-domain research c ommunity.
****

VisCUIT: Visual Auditor for Bias in CNN Image Classifier
Seongmin Lee, Zijie J. Wang, Judy Hoffman, Duen Horng (Polo) Chau; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 21475-21483
CNN image classifiers are widely used, thanks to their efficiency and accuracy. However, they can suffer from biases that impede their practical applications. M ost existing bias investigation techniques are either inapplicable to general im age classification tasks or require significant user efforts in perusing all dat a subgroups to manually specify which data attributes to inspect. We present Vis CUIT, an interactive visualization system that reveals how and why a CNN classif ier is biased. VisCUIT visually summarizes the subgroups on which the classifier underperforms and helps users discover and characterize the cause of the underp erformances by revealing image concepts responsible for activating neurons that contribute to misclassifications. VisCUIT runs in modern browsers and is open-so urce, allowing people to easily access and extend the tool to other model archit ectures and datasets. VisCUIT is available at the following public demo link: ht tps://poloclub.github.io/VisCUIT. A video demo is available at https://youtu.be/

eNDbSyM4R_4.
****
Automatic Synthesis of Diverse Weak Supervision Sources for Behavior Analysis
Albert Tseng, Jennifer J. Sun, Yisong Yue; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2211-2220
Obtaining annotations for large training sets is expensive, especially in settin gs where domain knowledge is required, such as behavior analysis. Weak supervisi on has been studied to reduce annotation costs by using weak labels from task-sp ecific labeling functions (LFs) to augment ground truth labels. However, domain experts still need to hand-craft different LFs for different tasks, limiting sca lability. To reduce expert effort, we present AutoSWAP: a framework for automati cally synthesizing data-efficient task-level LFs. The key to our approach is to efficiently represent expert knowledge in a reusable domain-specific language an d more general domain-level LFs, with which we use state-of-the-art program synt hesis techniques and a small labeled dataset to generate task-level LFs. Additio nally, we propose a novel structural diversity cost that allows for efficient sy nthesis of diverse sets of LFs, further improving AutoSWAP's performance. We eva luate AutoSWAP in three behavior analysis domains and demonstrate that AutoSWAP outperforms existing approaches using only a fraction of the data. Our results s uggest that AutoSWAP is an effective way to automatically generate LFs that can significantly reduce expert effort for behavior analysis.
****
Transferability Estimation Using Bhattacharyya Class Separability
Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, Thomas Mens ink; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2022, pp. 9172-9182
Transfer learning has become a popular method for leveraging pre-trained models in computer vision. However, without performing computationally expensive fine-t uning, it is difficult to quantify which pre-trained source models are suitable for a specific target task, or, conversely, to which tasks a pre-trained source model can be easily adapted to. In this work, we propose Gaussian Bhattacharyya Coefficient (GBC), a novel method for quantifying transferability between a sour ce model and a target dataset. In a first step we embed all target images in the feature space defined by the source model, and represent them with per-class Ga ussians. Then, we estimate their pairwise class separability using the Bhattacha ryya coefficient, yielding a simple and effective measure of how well the source model transfers to the target task. We evaluate GBC on image classification tas ks in the context of dataset and architecture selection. Further, we also perfor m experiments on the more complex semantic segmentation transferability estimati on task. We demonstrate that GBC outperforms state-of-the-art transferability me trics on most evaluation criteria in the semantic segmentation settings, matches the performance of top methods for dataset transferability in image classificat ion, and performs best on architecture selection problems for image classificati on.
****
DirecFormer: A Directed Attention in Transformer Approach to Robust Action Recog nition
Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 20030-20040
Human action recognition has recently become one ofthe popular research topics i n the computer vision community. Various 3D-CNN based methods have been presente d to tackle both the spatial and temporal dimensions in thetask of video action recognition with competitive results.However, these methods have suffered some f undamentallimitations such as lack of robustness and generalization,e.g., how do es the temporal ordering of video frames af-fect the recognition results? This w ork presents a novelend-to-end Transformer-based Directed Attention (Direc-Forme r) framework1for robust action recognition. The method takes a simple but novel perspective of Transformer-based approach to understand the right order of seque nce actions. Therefore, the contributions of this work are three-fold. Firstly,

we introduce the problem of ordered temporal learning issues to the action recog nition problem. Secondly, a new Directed Attention mechanism is introduced to un derstand and provide attentions to human actions in the right order. Thirdly, we introduce the conditional dependency in action sequence modeling that includes orders and classes. The proposed approach consistently achieves the state-of-the -art (SOTA) results compared with the recent action recognition methods [4, 15, 62, 64], on three standard large-scale benchmarks, i.e. Jester, Kinetics-400 and Something-Something-V2.

****

Hierarchical Self-Supervised Representation Learning for Movie Understanding
Fanyi Xiao, Kaustav Kundu, Joseph Tighe, Davide Modolo; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9727 -9736

Most self-supervised video representation learning approaches focus on action re cognition. In contrast, in this paper we focus on self-supervised video learning for movie understanding and propose a novel hierarchical self-supervised pretra ining strategy that separately pretrains each level of our hierarchical movie un derstanding model. Specifically, we propose to pretrain the low-level video back bone using a contrastive learning objective, while pretrain the higher-level vid eo contextualizer using an event mask prediction task, which enables the usage o f different data sources for pretraining different levels of the hierarchy. We f irst show that our self-supervised pretraining strategies are effective and lead to improved performance on all tasks and metrics on VidSitu benchmark (e.g., im proving on semantic role prediction from 47% to 61% CIDEr scores). We further de monstrate the effectiveness of our contextualized event features on LVU tasks, b oth when used alone and when combined with instance features, showing their comp lementarity.

****

Robust Egocentric Photo-Realistic Facial Expression Transfer for Virtual Reality
Amin Jourabloo, Fernando De la Torre, Jason Saragih, Shih-En Wei, Stephen Lombar di, Te-Li Wang, Danielle Belko, Autumn Trimble, Hernan Badino; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 20323-20332

Social presence, the feeling of being there with a "real" person, will fuel the next generation of communication systems driven by digital humans in virtual rea lity (VR). The best 3D video-realistic VR avatars that minimize the uncanny effe ct rely on person-specific (PS) models. However, these PS models are time-consum ing to build and are typically trained with limited data variability, which resu lts in poor generalization and robustness. Major sources of variability that aff ects the accuracy of facial expression transfer algorithms include using differe nt VR headsets (e.g., camera configuration, slop of the headset), facial appeara nce changes over time (e.g., beard, make-up), and environmental factors (e.g., l ighting, backgrounds). This is a major drawback for the scalability of these mod els in VR. This paper makes progress in overcoming these limitations by proposin g an end-to-end multi-identity architecture (MIA) trained with specialized augme ntation strategies. MIA drives the shape component of the avatar from three came ras in the VR headset (two eyes, one mouth), in untrained subjects, using minima l personalized information (i.e., neutral 3D mesh shape). Similarly, if the PS t exture decoder is available, MIA is able to drive the full avatar (shape+texture) robustly outperforming PS models in challenging scenarios. Our key contributio n to improve robustness and generalization, is that our method implicitly decoup les, in an unsupervised manner, the facial expression from nuisance factors (e.g ., headset, environment, facial appearance). We demonstrate the superior perform ance and robustness of the proposed method versus state-of-the-art PS approaches in a variety of experiments.

****

Does Robustness on ImageNet Transfer to Downstream Tasks?
Yutaro Yamada, Mayu Otani; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 9215-9224

As clean ImageNet accuracy nears its ceiling, the research community is increasi

ngly more concerned about robust accuracy under distributional shifts. While a v ariety of methods have been proposed to robustify neural networks, these techniq ues often target models trained on ImageNet classification. At the same time, it is a common practice to use ImageNet pretrained backbones for downstream tasks such as object detection, semantic segmentation, and image classification from d ifferent domains. This raises a question: Can these robust image classifiers tra nsfer robustness to downstream tasks? For object detection and semantic segmenta tion, we find that a vanilla Swin Transformer, a variant of Vision Transformer t ailored for dense prediction tasks, transfers robustness better than Convolution al Neural Networks that are trained to be robust to the corrupted version of Ima geNet. For CIFAR10 classification, we find that models that are robustified for ImageNet do not retain robustness when fully fine-tuned. These findings suggest that current robustification techniques tend to emphasize ImageNet evaluations. Moreover, network architecture is a strong source of robustness when we consider transfer learning.

****

Propagation Regularizer for Semi-Supervised Learning With Extremely Scarce Label ed Samples

Noo-ri Kim, Jee-Hyong Lee; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 14401-14410

Semi-supervised learning (SSL) is a method to make better models using a large n umber of easily accessible unlabeled data along with a small number of labeled d ata obtained at a high cost. Most of existing SSL studies focus on the cases whe re sufficient amount of labeled samples are available, tens to hundreds labeled samples for each class, which still requires a lot of labeling cost. In this pap er, we focus on SSL environment with extremely scarce labeled samples, only 1 or 2 labeled samples per class, where most of existing methods fail to learn. We p ropose a propagation regularizer which can achieve efficient and effective learn ing with extremely scarce labeled samples by suppressing confirmation bias. In a ddition, for the realistic model selection in the absence of the validation data set, we also propose a model selection method based on our propagation regulariz er. The proposed methods show 70.9%, 30.3%, and 78.9% accuracy on CIFAR-10, CIFA R-100, SVHN dataset with just one labeled sample per class, which are improved b y 8.9% to 120.2% compared to the existing approaches. And our proposed methods a lso show good performance on a higher resolution dataset, STL-10.

****

Bailando: 3D Dance Generation by Actor-Critic GPT With Choreographic Memory

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2022, pp. 11050-11059

Driving 3D characters to dance following a piece of music is highly challenging due to the spatial constraints applied to poses by choreography norms. In additi on, the generated dance sequence also needs to maintain temporal coherency with different music genres. To tackle these challenges, we propose a novel music-to-dance framework, Bailando, with two powerful components: 1) a choreographic memo ry that learns to summarize meaningful dancing units from 3D pose sequence to a quantized codebook, 2) an actor-critic Generative Pre-trained Transformer (GPT) that composes these units to a fluent dance coherent to the music. With the lear ned choreographic memory, dance generation is realized on the quantized units th at meet high choreography standards, such that the generated dancing sequences a re confined within the spatial constraints. To achieve synchronized alignment be tween diverse motion tempos and music beats, we introduce an actor-critic-based reinforcement learning scheme to the GPT with a newly-designed beat-align reward function. Extensive experiments on the standard benchmark demonstrate that our proposed framework achieves state-of-the-art performance both qualitatively and quantitatively. Notably, the learned choreographic memory is shown to discover h uman-interpretable dancing-style poses in an unsupervised manner.

****

Faithful Extreme Rescaling via Generative Prior Reciprocated Invertible Represen tations

Zhixuan Zhong, Liangyu Chai, Yang Zhou, Bailin Deng, Jia Pan, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5708-5717

This paper presents a Generative prior ReciprocAted Invertible rescaling Network (GRAIN) for generating faithful high-resolution (HR) images from low-resolution (LR) invertible images with an extreme upscaling factor (64x). Previous researches have leveraged the prior knowledge of a pretrained GAN model to generate high-quality upscaling results. However, they fail to produce pixel-accurate results due to the highly ambiguous extreme mapping process. We remedy this problem by introducing a reciprocated invertible image rescaling process, in which high-resolution information can be delicately embedded into an invertible low-resolution image and generative prior for a faithful HR reconstruction. In particular, the invertible LR features not only carry significant HR semantics, but also are trained to predict scale-specific latent codes, yielding a preferable utilization of generative features. On the other hand, the enhanced generative prior is re-injected to the rescaling process, compensating the lost details of the invertible rescaling. Our reciprocal mechanism perfectly integrates the advantages of invertible encoding and generative prior, leading to the first feasible extreme rescaling solution. Extensive experiments demonstrate superior performance against state-of-the-art upscaling methods. Code is available at https://github.com/cszzx/GRAIN.

****

Distillation Using Oracle Queries for Transformer-Based Human-Object Interaction Detection

Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19558-19567

Transformer-based methods have achieved great success in the field of human-object interaction (HOI) detection. However, these models tend to adopt semantically ambiguous queries, which lowers the transformer's representation learning power. Moreover, there are a very limited number of labeled human-object pairs for most images in existing datasets, which constrains the transformer's set prediction power. To handle the first problem, we propose an efficient knowledge distillation model, named Distillation using Oracle Queries (DOQ), which shares parameters between teacher and student networks. The teacher network adopts oracle queries that are semantically clear and generates high-quality decoder embeddings. By mimicking both the attention maps and decoder embeddings of the teacher network, the representation learning power of the student network is significantly promoted. To address the second problem, we introduce an efficient data augmentation method, named Context-Consistent Stitching (CCS), which generates complicated images online. Each new image is obtained by stitching labeled human-object pairs cropped from multiple training images. By selecting source images with similar context, the new synthesized image is made visually realistic. Our methods significantly promote both the accuracy and training efficiency of transformer-based HOI detection models. Experimental results show that our proposed approach consistently outperforms state-of-the-art methods on three benchmarks: HICO-DET, HOI-A, and V-COCO. Code will be released soon.

****

Proto2Proto: Can You Recognize the Car, the Way I Do?

Monish Keswani, Sriranjani Ramakrishnan, Nishant Reddy, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10233-10243

Prototypical methods have recently gained a lot of attention due to their intrinsic interpretable nature, which is obtained through the prototypes. With growing use cases of model reuse and distillation, there is a need to also study transfer of interpretability from one model to another. We present Proto2Proto, a novel method to transfer interpretability of one prototypical part network to another via knowledge distillation. Our approach aims to add interpretability to the "dark" knowledge transferred from the teacher to the shallower student model. We propose two novel losses: "Global Explanation" loss and "Patch-Prototype Corresp

ondence" loss to facilitate such a transfer. Global Explanation loss forces the student prototypes to be close to teacher prototypes, and Patch-Prototype Correspondence loss enforces the local representations of the student to be similar to that of the teacher. Further, we propose three novel metrics to evaluate the student's proximity to the teacher as measures of interpretability transfer in our settings. We qualitatively and quantitatively demonstrate the effectiveness of our method on CUB-200-2011 and Stanford Cars datasets. Our experiments show that the proposed method indeed achieves interpretability transfer from teacher to student while simultaneously exhibiting competitive performance. The code is available at https://github.com/archmaester/proto2proto

****

Learning Local-Global Contextual Adaptation for Multi-Person Pose Estimation

Nan Xue, Tianfu Wu, Gui-Song Xia, Liangpei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13065-13074

This paper studies the problem of multi-person pose estimation in a bottom-up fashion. With a new and strong observation that the localization issue of the center-offset formulation can be remedied in a local-window search scheme in an ideal situation, we propose a multi-person pose estimation approach, dubbed as LOGO-CAP, by learning the LOcal-GlObal Contextual Adaptation for human Pose. Specifically, our approach learns the keypoint attraction maps (KAMs) from the local keypoints expansion maps (KEMs) in small local windows in the first step, which are subsequently treated as dynamic convolutional kernels on the keypoints-focused global heatmaps for contextual adaptation, achieving accurate multi-person pose estimation. Our method is end-to-end trainable with near real-time inference speed in a single forward pass, obtaining state-of-the-art performance on the COCO keypoint benchmark for bottom-up human pose estimation. With the COCO trained model, our method also outperforms prior arts by a large margin on the challenging OCHuman dataset.

****

Learning Video Representations of Human Motion From Synthetic Data

Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, Qin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20197-20207

In this paper, we take an early step towards video representation learning of human actions with the help of largescale synthetic videos, particularly for human motion representation enhancement. Specifically, we first introduce an automatic action-related video synthesis pipeline based on a photorealistic video game. A large-scale human action dataset named GATA (GTA Animation Transformed Actions) is then built by the proposed pipeline, which includes 8.1 million action clips spanning over 28K action classes. Based on the presented dataset, we design a contrastive learning framework for human motion representation learning, which shows significant performance improvements on several typical video datasets for action recognition, e.g., Charades, HAA 500 and NTU-RGB. Besides, we further explore a domain adaptation method based on cross-domain positive pairs mining to alleviate the domain gap between synthetic and realistic data. Extensive properties analyses of learned representation are conducted to demonstrate the effectiveness of the proposed dataset for enhancing human motion representation learning.

****

TVConv: Efficient Translation Variant Convolution for Layout-Aware Visual Processing

Jierun Chen, Tianlang He, Weipeng Zhuo, Li Ma, Sangtae Ha, S.-H. Gary Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12548-12558

As convolution has empowered many smart applications, dynamic convolution further equips it with the ability to adapt to diverse inputs. However, the static and dynamic convolutions are either layout-agnostic or computation-heavy, making it inappropriate for layout-specific applications, e.g., face recognition and medical image segmentation. We observe that these applications naturally exhibit the characteristics of large intra-image (spatial) variance and small cross-image v

ariance. This observation motivates our efficient translation variant convolutio n (TVConv) for layout-aware visual processing. Technically, TVConv is composed o f affinity maps and a weight-generating block. While affinity maps depict pixel-paired relationships gracefully, the weight-generating block can be explicitly o verparameterized for better training while maintaining efficient inference. Alth ough conceptually simple, TVConv significantly improves the efficiency of the co nvolution and can be readily plugged into various network architectures. Extensi ve experiments on face recognition show that TVConv reduces the computational co st by up to 3.1x and improves the corresponding throughput by 2.3x while maintai ning a high accuracy compared to the depthwise convolution. Moreover, for the sa me computation cost, we boost the mean accuracy by up to 4.21%. We also conduct experiments on the optic disc/cup segmentation task and obtain better generaliza tion performance, which helps mitigate the critical data scarcity issue. Code is available at https://github.com/JierunChen/TVConv.

****

Dual Adversarial Adaptation for Cross-Device Real-World Image Super-Resolution
Xiaoqian Xu, Pengxu Wei, Weikai Chen, Yang Liu, Mingzhi Mao, Liang Lin, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 5667-5676

Due to the sophisticated imaging process, an identical scene captured by differe nt cameras could exhibit distinct imaging patterns, introducing distinct profici ency among the super-resolution (SR) models trained on images from different dev ices. In this paper, we investigate a novel and practical task coded cross-devic e SR, which strives to adapt a real-world SR model trained on the paired images captured by one camera to low-resolution (LR) images captured by arbitrary targe t devices. The proposed task is highly challenging due to the absence of paired data from various imaging devices. To address this issue, we propose an unsuperv ised domain adaptation mechanism for real-world SR, named Dual ADversarial Adapt ation (DADA), which only requires LR images in the target domain with available real paired data from a source camera. DADA employs the Domain-Invariant Attenti on (DIA) module to establish the basis of target model training even without HR supervision. Furthermore, the dual framework of DADA facilitates an Inter-domain Adversarial Adaptation (InterAA) in one branch for two LR input images from two domains, and an Intra-domain Adversarial Adaptation (IntraAA) in two branches f or an LR input image. InterAA and IntraAA together improve the model transferabi lity from the source domain to the target. We empirically conduct experiments un der six Real to Real adaptation settings among three different cameras, and achi eve superior performance compared with existing state-of-the-art approaches. We also evaluate the proposed DADA to address the adaptation to the video camera, w hich presents a promising research topic to promote the wide applications of rea l-world super-resolution. Our source code is publicly available at https://githu b.com/lonelyhope/DADA.git.

****

FS6D: Few-Shot 6D Pose Estimation of Novel Objects
Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, Qifeng Chen; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6 814-6824

6D object pose estimation networks are limited in their capability to scale to l arge numbers of object instances due to the close-set assumption and their relia nce on high-fidelity object CAD models. In this work, we study a new open set pr oblem; the few-shot 6D object poses estimation: estimating the 6D pose of an unk nown object by a few support views without extra training. To tackle the problem , we point out the importance of fully exploring the appearance and geometric re lationship between the given support views and query scene patches and propose a dense prototypes matching framework by extracting and matching dense RGBD proto types with transformers. Moreover, we show that the priors from diverse appearan ces and shapes are crucial to the generalization capability under the problem se tting and thus propose a large-scale RGBD photorealistic dataset (ShapeNet6D) fo r network pre-training. A simple and effective online texture blending approach is also introduced to eliminate the domain gap from the synthesis dataset, which

enriches appearance diversity at a low cost. Finally, we discuss possible solutions to this problem and establish benchmarks on popular datasets to facilitate future research.
****

Habitat-Web: Learning Embodied Object-Search Strategies From Human Demonstrations at Scale
Ram Ramrakhya, Eric Undersander, Dhruv Batra, Abhishek Das; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5173-5183
We present a large-scale study of imitating human demonstrations on tasks that require a virtual robot to search for objects in new environments - (1) ObjectGoal Navigation (e.g. 'find & go to a chair') and (2) Pick&Place (e.g. 'find mug, pick mug, find counter, place mug on counter'). First, we develop a virtual teleoperation data-collection infrastructure - connecting Habitat simulator running in a web browser to Amazon Mechanical Turk, allowing remote users to teleoperate virtual robots, safely and at scale. We collect 80k demonstrations for ObjectNav and 12k demonstrations for Pick&Place, which is an order of magnitude larger than existing human demonstration datasets in simulation or on real robots. Our virtual teleoperation data contains 29.3M actions, and is equivalent to 22.6k hours of real-world teleoperation time, and illustrates rich, diverse strategies for solving the tasks. Second, we use this data to answer the question - how does large-scale imitation learning (IL) (which has not been hitherto possible) compare to reinforcement learning (RL) (which is the status quo)? On ObjectNav, we find that IL (with no bells or whistles) using 70k human demonstrations outperforms RL using 240k agent-gathered trajectories. This effectively establishes an 'exchange rate' - a single human demonstration appears to be worth 4 agent-gathered ones. More importantly, we find the IL-trained agent learns efficient object-search behavior from humans - it peeks into rooms, checks corners for small objects, turns in place to get a panoramic view - none of these are exhibited as prominently by the RL agent, and to induce these behaviors via contemporary RL techniques would require tedious reward engineering. Finally, accuracy vs. training data size plots show promising scaling behavior, suggesting that simply collecting more demonstrations is likely to advance the state of art further. On Pick&Place, the comparison is starker - IL agents achieve 18% success on episodes with new object-receptacle locations when trained with 9.5k human demonstrations, while RL agents fail to get beyond 0%. Overall, our work provides compelling evidence for investing in large-scale imitation learning. Project page: https://ram81.github.io/projects/habitat-web.
****

The Probabilistic Normal Epipolar Constraint for Frame-to-Frame Rotation Optimization Under Uncertain Feature Positions
Dominik Muhle, Lukas Koestler, Nikolaus Demmel, Florian Bernard, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1819-1828
The estimation of the relative pose of two camera views is a fundamental problem in computer vision. Kneip et al. proposed to solve this problem by introducing the normal epipolar constraint (NEC). However, their approach does not take into account uncertainties, so that the accuracy of the estimated relative pose is highly dependent on accurate feature positions in the target frame. In this work, we introduce the probabilistic normal epipolar constraint (PNEC) that overcomes this limitation by accounting for anisotropic and inhomogeneous uncertainties in the feature positions. To this end, we propose a novel objective function, along with an efficient optimization scheme that effectively minimizes our objective while maintaining real-time performance. In experiments on synthetic data, we demonstrate that the novel PNEC yields more accurate rotation estimates than the original NEC and several popular relative rotation estimation algorithms. Furthermore, we integrate the proposed method into a state-of-the-art monocular rotation-only odometry system and achieve consistently improved results for the real-world KITTI dataset.
****

Vision-Language Pre-Training for Boosting Scene Text Detectors
Sibo Song, Jianqiang Wan, Zhibo Yang, Jun Tang, Wenqing Cheng, Xiang Bai, Cong Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15681-15691
Recently, vision-language joint representation learning has proven to be highly effective in various scenarios. In this paper, we specifically adapt vision-language joint learning for scene text detection, a task that intrinsically involves cross-modal interaction between the two modalities: vision and language, since text is the written form of language. Concretely, we propose to learn contextualized, joint representations through vision-language pre-training, for the sake of enhancing the performance of scene text detectors. Towards this end, we devise a pre-training architecture with an image encoder, a text encoder and a cross-modal encoder, as well as three pretext tasks: image-text contrastive learning (ITC), masked language modeling (MLM) and word-in-image prediction (WIP). The pre-trained model is able to produce more informative representations with richer semantics, which could readily benefit existing scene text detectors (such as EAST and PSENet) in the down-stream text detection task. Extensive experiments on standard benchmarks demonstrate that the proposed paradigm can significantly improve the performance of various representative text detectors, outperforming previous pre-training approaches. The code and pre-trained models will be publicly released.
****
Reflection and Rotation Symmetry Detection via Equivariant Learning
Ahyun Seo, Byungjin Kim, Suha Kwak, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9539-9548
The inherent challenge of detecting symmetries stems from arbitrary orientations of symmetry patterns; a reflection symmetry mirrors itself against an axis with a specific orientation while a rotation symmetry matches its rotated copy with a specific orientation. Discovering such symmetry patterns from an image thus benefits from an equivariant feature representation, which varies consistently with reflection and rotation of the image. In this work, we introduce a group-equivariant convolutional network for symmetry detection, dubbed EquiSym, which leverages equivariant feature maps with respect to a dihedral group of reflection and rotation. The proposed network is built end-to-end with dihedrally-equivariant layers and trained to output a spatial map for reflection axes or rotation centers. We also present a new dataset, DENse and DIverse symmetry (DENDI), which mitigates limitations of existing benchmarks for reflection and rotation symmetry detection. Experiments show that our method achieves the state of the arts in symmetry detection on LDRS and DENDI datasets.
****
BoostMIS: Boosting Medical Image Semi-Supervised Learning With Adaptive Pseudo Labeling and Informative Active Annotation
Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, Beng Chin Ooi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20666-20676
In this paper, we propose a novel semi-supervised learning (SSL) framework named BoostMIS that combines adaptive pseudo labeling and informative active annotation to unleash the potential of medical image SSL models: (1) BoostMIS can adaptively leverage the cluster assumption and consistency regularization of the unlabeled data according to the current learning status. This strategy can adaptively generate one-hot "hard" labels converted from task model predictions for better task model training. (2) For the unselected unlabeled images with low confidence, we introduce an Active learning (AL) algorithm to find the informative samples as the annotation candidates by exploiting virtual adversarial perturbation and model's density-aware entropy. These informative candidates are subsequently fed into the next training cycle for better SSL label propagation. Notably, the adaptive pseudo-labeling and informative active annotation form a learning closed-loop that are mutually collaborative to boost medical image SSL. To verify the effectiveness of the proposed method, we collected a metastatic epidural spinal cord compression (MESCC) dataset that aims to optimize MESCC diagnosis and class

ification for improved specialist referral and treatment. We conducted an extens ive experimental study of BoostMIS on MESCC and another public dataset COVIDx. T he experimental results verify our framework's effectiveness and generalisabilit y for different medical image datasets with a significant improvement over vario us state-of-the-art methods.
****

Simple but Effective: CLIP Embeddings for Embodied AI
Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, Aniruddha Kembhavi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14829-14838
Contrastive language image pretraining (CLIP) encoders have been shown to be beneficial for a range of visual tasks from classification and detection to captioning and image manipulation. We investigate the effectiveness of CLIP visual backbones for Embodied AI tasks. We build incredibly simple baselines, named EmbCLIP, with no task specific architectures, inductive biases (such as the use of semantic maps), auxiliary tasks during training, or depth maps--yet we find that our improved baselines perform very well across a range of tasks and simulators. EmbCLIP tops the RoboTHOR ObjectNav leaderboard by a huge margin of 20 pts (Success Rate). It tops the iTHOR 1-Phase Rearrangement leaderboard, beating the next best submission, which employs Active Neural Mapping, and more than doubling the % Fixed Strict metric (0.08 to 0.17). It also beats the winners of the 2021 Habitat ObjectNav Challenge, which employ auxiliary tasks, depth maps, and human demonstrations, and those of the 2019 Habitat PointNav Challenge. We evaluate the ability of CLIP's visual representations at capturing semantic information about input observations--primitives that are useful for navigation-heavy embodied tasks--and find that CLIP's representations encode these primitives more effectively than ImageNet-pretrained backbones. Finally, we extend one of our baselines, producing an agent capable of zero-shot object navigation that can navigate to objects that were not used as targets during training. Our code and models are available at https://github.com/allenai/embodied-clip.
****

NomMer: Nominate Synergistic Context in Vision Transformer for Visual Recognition
Hao Liu, Xinghua Jiang, Xin Li, Zhimin Bao, Deqiang Jiang, Bo Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12073-12082
Recently, Vision Transformers (ViT), with the self-attention (SA) as the de facto ingredients, have demonstrated great potential in the computer vision community. For the sake of trade-off between efficiency and performance, a group of works merely perform SA operation within local patches, whereas the global contextual information is abandoned, which would be indispensable for visual recognition tasks. To solve the issue, the subsequent global-local ViTs take a stab at marrying local SA with global one in parallel or alternative way in the model. Nevertheless, the exhaustively combined local and global context may exist redundancy for various visual data, and the receptive field within each layer is fixed. Alternatively, a more graceful way is that global and local context can adaptively contribute per se to accommodate different visual data. To achieve this goal, we in this paper propose a novel ViT architecture, termed NomMer, which can dynamically Nominate the synergistic global-local context in vision transforMer. By investigating the working pattern of our proposed NomMer, we further explore what context information is focused. Beneficial from this "dynamic nomination" mechanism, without bells and whistles, the NomMer can not only achieve 84.5% Top-1 classification accuracy on ImageNet with only 73M parameters, but also show promising performance on dense prediction tasks, i.e., object detection and semantic segmentation. The code and models are publicly available at https://github.com/TencentYoutuResearch/VisualRecognition-NomMer.
****

HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction
Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, Li Yi; Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR), 2022, pp. 21013-21022
We present HOI4D, a large-scale 4D egocentric dataset with rich annotations, to catalyze the research of category-level human-object interaction. HOI4D consists of 2.4M RGB-D egocentric video frames over 4000 sequences collected by 9 participants interacting with 800 different object instances from 16 categories over 610 different indoor rooms. Frame-wise annotations for panoptic segmentation, motion segmentation, 3D hand pose, category-level object pose and hand action have also been provided, together with reconstructed object meshes and scene point clouds. With HOI4D, we establish three benchmarking tasks to promote category-level HOI from 4D visual signals including semantic segmentation of 4D dynamic point cloud sequences, category-level object pose tracking, and egocentric action segmentation with diverse interaction targets. In-depth analysis shows HOI4D poses great challenges to existing methods and produces huge research opportunities.
****

Collaborative Transformers for Grounded Situation Recognition
Junhyeong Cho, Youngseok Yoon, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19659-19668
Grounded situation recognition is the task of predicting the main activity, entities playing certain roles within the activity, and bounding-box groundings of the entities in the given image. To effectively deal with this challenging task, we introduce a novel approach where the two processes for activity classification and entity estimation are interactive and complementary. To implement this idea, we propose Collaborative Glance-Gaze TransFormer (CoFormer) that consists of two modules: Glance transformer for activity classification and Gaze transformer for entity estimation. Glance transformer predicts the main activity with the help of Gaze transformer that analyzes entities and their relations, while Gaze transformer estimates the grounded entities by focusing only on the entities relevant to the activity predicted by Glance transformer. Our CoFormer achieves the state of the art in all evaluation metrics on the SWiG dataset. Training code and model weights are available at https://github.com/jhcho99/CoFormer.
****

DyRep: Bootstrapping Training With Dynamic Re-Parameterization
Tao Huang, Shan You, Bohan Zhang, Yuxuan Du, Fei Wang, Chen Qian, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 588-597
Structural re-parameterization (Rep) methods achieve noticeable improvements on simple VGG-style networks. Despite the prevalence, current Rep methods simply re-parameterize all operations into an augmented network, including those that rarely contribute to the model's performance. As such, the price to pay is an expensive computational overhead to manipulate these unnecessary behaviors. To eliminate the above caveats, we aim to bootstrap the training with minimal cost by devising a dynamic re-parameterization (DyRep) method, which encodes Rep technique into the training process that dynamically evolves the network structures. Concretely, our proposal adaptively finds the operations which contribute most to the loss in the network, and applies Rep to enhance their representational capacity. Besides, to suppress the noisy and redundant operations introduced by Rep, we devise a de-parameterization technique for a more compact re-parameterization. With this regard, DyRep is more efficient than Rep since it smoothly evolves the given network instead of constructing an over-parameterized network. Experimental results demonstrate our effectiveness, e.g., DyRep improves the accuracy of ResNet-18 by 2.04% on ImageNet and reduces 22% runtime over the baseline.
****

Not All Labels Are Equal: Rationalizing the Labeling Costs for Training Object Detection
Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixé, Jose M. Alvarez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14492-14501
Deep neural networks have reached high accuracy on object detection but their success hinges on large amounts of labeled data. To reduce the labels dependency, various active learning strategies have been proposed, typically based on the co

nfidence of the detector. However, these methods are biased towards high-perform ing classes and can lead to acquired datasets that are not good representatives of the testing set data. In this work, we propose a unified framework for active learning, that considers both the uncertainty and the robustness of the detecto r, ensuring that the network performs well in all classes. Furthermore, our meth od leverages auto-labeling to suppress a potential distribution drift while boos ting the performance of the model. Experiments on PASCAL VOC07+12 and MS-COCO sh ow that our method consistently outperforms a wide range of active learning meth ods, yielding up to a 7.7% improvement in mAP, or up to 82% reduction in labelin g cost. Code will be released upon acceptance of the paper.
****

CPPF: Towards Robust Category-Level 9D Pose Estimation in the Wild
Yang You, Ruoxi Shi, Weiming Wang, Cewu Lu; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6866-6875
In this paper, we tackle the problem of category-level 9D pose estimation in the wild, given a single RGB-D frame. Using supervised data of real-world 9D poses is tedious and erroneous, and also fails to generalize to unseen scenarios. Besi des, category-level pose estimation requires a method to be able to generalize t o unseen objects at test time, which is also challenging. Drawing inspirations f rom traditional point pair features (PPFs), in this paper, we design a novel Cat egory-level PPF (CPPF) voting method to achieve accurate, robust and generalizab le 9D pose estimation in the wild. To obtain robust pose estimation, we sample n umerous point pairs on an object, and for each pair our model predicts necessary SE(3)-invariant voting statistics on object centers, orientations and scales. A novel coarse-to-fine voting algorithm is proposed to eliminate noisy point pair samples and generate final predictions from the population. To get rid of false positives in the orientation voting process, an auxiliary binary disambiguating classification task is introduced for each sampled point pair. In order to dete ct objects in the wild, we carefully design our sim-to-real pipeline by training on synthetic point clouds only, unless objects have ambiguous poses in geometry . Under this circumstance, color information is leveraged to disambiguate these poses. Results on standard benchmarks show that our method is on par with curren t state of the arts with real-world training data. Extensive experiments further show that our method is robust to noise and gives promising results under extre mely challenging scenarios. Our code is available on https://github.com/qq456cvb /CPPF.
****

Interact Before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Acti on Recognition
Lijin Yang, Yifei Huang, Yusuke Sugano, Yoichi Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14722-1 4732
Unsupervised domain adaptive video action recognition aims to recognize actions of a target domain using a model trained with only out-of-domain (source) annota tions. The inherent complexity of videos makes this task challenging but also pr ovides ground for leveraging multi-modal inputs (e.g., RGB, Flow, Audio). Most p revious works utilize the multi-modal information by either aligning each modali ty individually or learning representation via cross-modal self-supervision. Dif ferent from previous works, we find that the cross-domain alignment can be more effectively done by using cross-modal interaction first. Cross-modal knowledge i nteraction allows other modalities to supplement missing transferable informatio n because of the cross-modal complementarity. Also, the most transferable aspect s of data can be highlighted using cross-modal consensus. In this work, we prese nt a novel model that jointly considers these two characteristics for domain ada ptive action recognition. We achieve this by implementing two modules, where the first module exchanges complementary transferable information across modalities through the semantic space, and the second module finds the most transferable s patial region based on the consensus of all modalities. Extensive experiments va lidate that our proposed method can significantly outperform the state-of-the-ar t methods on multiple benchmark datasets, including the complex fine-grained dat

aset EPIC-Kitchens-100.
****

Interactive Disentanglement: Learning Concepts by Interacting With Their Prototype Representations
Wolfgang Stammer, Marius Memmel, Patrick Schramowski, Kristian Kersting; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10317-10328
Learning visual concepts from raw images without strong supervision is a challenging task. In this work, we show the advantages of prototype representations for understanding and revising the latent space of neural concept learners. For this purpose, we introduce interactive Concept Swapping Networks (iCSNs), a novel framework for learning concept-grounded representations via weak supervision and implicit prototype representations. iCSNs learn to bind conceptual information to specific prototype slots by swapping the latent representations of paired images. This semantically grounded and discrete latent space facilitates human understanding and human-machine interaction. We support this claim by conducting experiments on our novel data set "Elementary Concept Reasoning" (ECR), focusing on visual concepts shared by geometric objects.
****

CDGNet: Class Distribution Guided Network for Human Parsing
Kunliang Liu, Ouk Choi, Jianming Wang, Wonjun Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4473-4482
The objective of human parsing is to partition a human in an image into constituent parts. This task involves labeling each pixel of the human image according to the classes. Since the human body comprises hierarchically structured parts, each body part of an image can have its sole position distribution characteristic. Probably, a human head is less likely to be under the feet, and arms are more likely to be near the torso. Inspired by this observation, we make instance class distributions by accumulating the original human parsing label in the horizontal and vertical directions, which can be utilized as supervision signals. Using these horizontal and vertical class distribution labels, the network is guided to exploit the intrinsic position distribution of each class. We combine two guided features to form a spatial guidance map, which is then superimposed onto the baseline network by multiplication and concatenation to distinguish the human parts precisely. We conducted extensive experiments to demonstrate the effectiveness and superiority of our method on three well-known benchmarks: LIP, ATR, and CIHP databases.
****

Recall@k Surrogate Loss With Large Batches and Similarity Mixup
Yash Patel, Giorgos Tolias, Ji■í Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7502-7511
This work focuses on learning deep visual representation models for retrieval by exploring the interplay between a new loss function, the batch size, and a new regularization approach. Direct optimization, by gradient descent, of an evaluation metric, is not possible when it is non-differentiable, which is the case for recall in retrieval. A differentiable surrogate loss for the recall is proposed in this work. Using an implementation that sidesteps the hardware constraints of the GPU memory, the method trains with a very large batch size, which is essential for metrics computed on the entire retrieval database. It is assisted by an efficient mixup regularization approach that operates on pairwise scalar similarities and virtually increases the batch size further. The suggested method achieves state-of-the-art performance in several image retrieval benchmarks when used for deep metric learning. For instance-level recognition, the method outperforms similar approaches that train using an approximation of average precision.
****

Direct Voxel Grid Optimization: Super-Fast Convergence for Radiance Fields Reconstruction
Cheng Sun, Min Sun, Hwann-Tzong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5459-5469

We present a super-fast convergence approach to reconstructing the per-scene radiance field from a set of images that capture the scene with known poses. This task, which is often applied to novel view synthesis, is recently revolutionized by Neural Radiance Field (NeRF) for its state-of-the-art quality and flexibility. However, NeRF and its variants require a lengthy training time ranging from hours to days for a single scene. In contrast, our approach achieves NeRF-comparable quality and converges rapidly from scratch in less than 15 minutes with a single GPU. We adopt a representation consisting of a density voxel grid for scene geometry and a feature voxel grid with a shallow network for complex view-dependent appearance. Modeling with explicit and discretized volume representations is not new, but we propose two simple yet non-trivial techniques that contribute to fast convergence speed and high-quality output. First, we introduce the post-activation interpolation on voxel density, which is capable of producing sharp surfaces in lower grid resolution. Second, direct voxel density optimization is prone to suboptimal geometry solutions, so we robustify the optimization process by imposing several priors. Finally, evaluation on five inward-facing benchmarks shows that our method matches, if not surpasses, NeRF's quality, yet it only takes about 15 minutes to train from scratch for a new scene. We will make our code publicly available.

****

Continual Test-Time Domain Adaptation

Qin Wang, Olga Fink, Luc Van Gool, Dengxin Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7201-7211

Test-time domain adaptation aims to adapt a source pre-trained model to a target domain without using any source data. Existing works mainly consider the case where the target domain is static. However, real-world machine perception systems are running in non-stationary and continually changing environments where the target domain distribution can change over time. Existing methods, which are mostly based on self-training and entropy regularization, can suffer from these non-stationary environments. Due to the distribution shift over time in the target domain, pseudo-labels become unreliable. The noisy pseudo-labels can further lead to error accumulation and catastrophic forgetting. To tackle these issues, we propose a continual test-time adaptation approach (CoTTA) which comprises two parts. Firstly, we propose to reduce the error accumulation by using weight-averaged and augmentation-averaged predictions which are often more accurate. On the other hand, to avoid catastrophic forgetting, we propose to stochastically restore a small part of the neurons to the source pre-trained weights during each iteration to help preserve source knowledge in the long-term. The proposed method enables the long-term adaptation for all parameters in the network. CoTTA is easy to implement and can be readily incorporated in off-the-shelf pre-trained models. We demonstrate the effectiveness of our approach on four classification tasks and a segmentation task for continual test-time adaptation, on which we outperform existing methods. Our code is available at https://qin.ee/cotta.

****

URetinex-Net: Retinex-Based Deep Unfolding Network for Low-Light Image Enhancement

Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, Jianmin Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5901-5910

Retinex model-based methods have shown to be effective in layer-wise manipulation with well-designed priors for low-light image enhancement. However, the commonly used hand-crafted priors and optimization-driven solutions lead to the absence of adaptivity and efficiency. To address these issues, in this paper, we propose a Retinex-based deep unfolding network (URetinex-Net), which unfolds an optimization problem into a learnable network to decompose a low-light image into reflectance and illumination layers. By formulating the decomposition problem as an implicit priors regularized model, three learning-based modules are carefully designed, responsible for data-dependent initialization, high-efficient unfolding optimization, and user-specified illumination enhancement, respectively. Particularly, the proposed unfolding optimization module, introducing two networks to

adaptively fit implicit priors in data-driven manner, can realize noise suppression and details preservation for the final decomposition results. Extensive experiments on real-world low-light images qualitatively and quantitatively demonstrate the effectiveness and superiority of the proposed method over state-of-the-art methods.

****

Towards Multi-Domain Single Image Dehazing via Test-Time Training

Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, Keyan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5831-5840

Recent years have witnessed significant progress in the area of single image dehazing, thanks to the employment of deep neural networks and diverse datasets. Most of the existing methods perform well when the training and testing are conducted on a single dataset. However, they are not able to handle different types of hazy images using a dehazing model trained on a particular dataset. One possible remedy is to perform training on multiple datasets jointly. However, we observe that this training strategy tends to compromise the model performance on individual datasets. Motivated by this observation, we propose a test-time training method which leverages a helper network to assist the dehazing model in better adapting to a domain of interest. Specifically, during the test time, the helper network evaluates the quality of the dehazing results, then directs the dehazing network to improve the quality by adjusting its parameters via self-supervision. Nevertheless, the inclusion of the helper network does not automatically ensure the desired performance improvement. For this reason, a meta-learning approach is employed to make the objectives of the dehazing and helper networks consistent with each other. We demonstrate the effectiveness of the proposed method by providing extensive supporting experiments.

****

Vox2Cortex: Fast Explicit Reconstruction of Cortical Surfaces From 3D MRI Scans With Geometric Deep Neural Networks

Fabian Bongratz, Anne-Marie Rickmann, Sebastian Pölsterl, Christian Wachinger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20773-20783

The reconstruction of cortical surfaces from brain magnetic resonance imaging (MRI) scans is essential for quantitative analyses of cortical thickness and sulcal morphology. Although traditional and deep learning-based algorithmic pipelines exist for this purpose, they have two major drawbacks: lengthy runtimes of multiple hours (traditional) or intricate post-processing, such as mesh extraction and topology correction (deep learning-based). In this work, we address both of these issues and propose Vox2Cortex, a deep learning-based algorithm that directly yields topologically correct, three-dimensional meshes of the boundaries of the cortex. Vox2Cortex leverages convolutional and graph convolutional neural networks to deform an initial template to the densely folded geometry of the cortex represented by an input MRI scan. We show in extensive experiments on three brain MRI datasets that our meshes are as accurate as the ones reconstructed by state-of-the-art methods in the field, without the need for time- and resource-intensive post-processing. To accurately reconstruct the tightly folded cortex, we work with meshes containing about 168,000 vertices at test time, scaling deep explicit reconstruction methods to a new level.

****

Deep Safe Multi-View Clustering: Reducing the Risk of Clustering Performance Degradation Caused by View Increase

Huayi Tang, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 202-211

Multi-view clustering has been shown to boost clustering performance by effectively mining the complementary information from multiple views. However, we observe that learning from data with more views is not guaranteed to achieve better clustering performance than from data with fewer views. To address this issue, we propose a general deep learning based framework that is guaranteed to reduce the risk of performance degradation caused by view increase. Concretely, the model

is trained to simultaneously extract complementary information and discard the m
eaningless noise by automatically selecting features. These two learning procedu
res are incorporated into one unified framework by the proposed optimization obj
ective. In theory, the empirical clustering risk of the model is no higher than
learning from data before the view increase and data of the new increased single
 view. Also, the expected clustering risk of the model under divergence-based lo
ss is no higher than that with high probability. Comprehensive experiments on be
nchmark datasets demonstrate the effectiveness and superiority of the proposed f
ramework in achieving safe multi-view clustering.
****

Dynamic MLP for Fine-Grained Image Classification by Leveraging Geographical and
 Temporal Information

Lingfeng Yang, Xiang Li, Renjie Song, Borui Zhao, Juntian Tao, Shihao Zhou, Jiaj
un Liang, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2022, pp. 10945-10954

Fine-grained image classification is a challenging computer vision task where va
rious species share similar visual appearances, resulting in misclassification i
f merely based on visual clues. Therefore, it is helpful to leverage additional
information, e.g., the locations and dates for data shooting, which can be easil
y accessible but rarely exploited. In this paper, we first demonstrate that exis
ting multimodal methods fuse multiple features only on a single dimension, which
 essentially has insufficient help in feature discrimination. To fully explore t
he potential of multimodal information, we propose a dynamic MLP on top of the i
mage representation, which interacts with multimodal features at a higher and br
oader dimension. The dynamic MLP is an efficient structure parameterized by the
learned embeddings of variable locations and dates. It can be regarded as an ada
ptive nonlinear projection for generating more discriminative image representati
ons in visual tasks. To our best knowledge, it is the first attempt to explore t
he idea of dynamic networks to exploit multimodal information in fine-grained im
age classification tasks. Extensive experiments demonstrate the effectiveness of
 our method. The t-SNE algorithm visually indicates that our technique improves
the recognizability of image representations that are visually similar but with
different categories. Furthermore, among published works across multiple fine-gr
ained datasets, dynamic MLP consistently achieves SOTA results and takes third p
lace in the iNaturalist challenge at FGVC8. Code is available at https://github.
com/megvii-research/DynamicMLPForFinegrained.
****

HP-Capsule: Unsupervised Face Part Discovery by Hierarchical Parsing Capsule Net
work

Chang Yu, Xiangyu Zhu, Xiaomei Zhang, Zidu Wang, Zhaoxiang Zhang, Zhen Lei; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 4032-4041

Capsule networks are designed to present the objects by a set of parts and their
 relationships, which provide an insight into the procedure of visual perception
. Although recent works have shown the success of capsule networks on simple obj
ects like digits, the human faces with homologous structures, which are suitable
 for capsules to describe, have not been explored. In this paper, we propose a H
ierarchical Parsing Capsule Network (HP-Capsule) for unsupervised face subpart-p
art discovery. When browsing large-scale face images without labels, the network
 first encodes the frequently observed patterns with a set of explainable subpar
t capsules. Then, the subpart capsules are assembled into part-level capsules th
rough a Transformer-based Parsing Module (TPM) to learn the compositional relati
ons between them. During training, as the face hierarchy is progressively built
and refined, the part capsules adaptively encode the face parts with semantic co
nsistency. HP-Capsule extends the application of capsule networks from digits to
 human faces and takes a step forward to show how the neural networks understand
 homologous objects without human intervention. Besides, HP-Capsule gives unsupe
rvised face segmentation results by the covered regions of part capsules, enabli
ng qualitative and quantitative evaluation. Experiments on BP4D and Multi-PIE da
tasets show the effectiveness of our method.

****
ScanQA: 3D Question Answering for Spatial Scene Understanding
Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, Motoaki Kawanabe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19129-19139
We propose a new 3D spatial understanding task of 3D Question Answering (3D-QA). In the 3D-QA task, models receive visual information from the entire 3D scene of the rich RGB-D indoor scan and answer the given textual questions about the 3D scene. Unlike the 2D-question answering of VQA, the conventional 2D-QA models suffer from problems with spatial understanding of object alignment and directions and fail the object identification from the textual questions in 3D-QA. We propose a baseline model for 3D-QA, named ScanQA model, where the model learns a fused descriptor from 3D object proposals and encoded sentence embeddings. This learned descriptor correlates the language expressions with the underlying geometric features of the 3D scan and facilitates the regression of 3D bounding boxes to determine described objects in textual questions. We collected human-edited question-answer pairs with free-form answers that are grounded to 3D objects in each 3D scene. Our new ScanQA dataset contains over 40K question-answer pairs from the 800 indoor scenes drawn from the ScanNet dataset. To the best of our knowledge, ScanQA is the first large-scale effort to perform object-grounded question-answering in 3D environments.
****
MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-Based Visual Question Answering
Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5089-5098
Knowledge-based visual question answering requires the ability of associating external knowledge for open-ended cross-modal scene understanding. One limitation of existing solutions is that they capture relevant knowledge from text-only knowledge bases, which merely contain facts expressed by first-order predicates or language descriptions while lacking complex but indispensable multimodal knowledge for visual understanding. How to construct vision-relevant and explainable multimodal knowledge for the VQA scenario has been less studied. In this paper, we propose MuKEA to represent multimodal knowledge by an explicit triplet to correlate visual objects and fact answers with implicit relations. To bridge the heterogeneous gap, we propose three objective losses to learn the triplet representations from complementary views: embedding structure, topological relation and semantic space. By adopting a pre-training and fine-tuning learning strategy, both basic and domain-specific multimodal knowledge are progressively accumulated for answer prediction. We outperform the state-of-the-art by 3.35% and 6.08% respectively on two challenging knowledge-required datasets: OK-VQA and KRVQA. Experimental results prove the complementary benefits of the multimodal knowledge with existing knowledge bases and the advantages of our end-to-end framework over the existing pipeline methods. The code is available at https://github.com/AndersonStra/MuKEA.
****
Class-Incremental Learning by Knowledge Distillation With Adaptive Feature Consolidation
Minsoo Kang, Jaeyoo Park, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16071-16080
We present a novel class incremental learning approach based on deep neural networks, which continually learns new tasks with limited memory for storing examples in the previous tasks. Our algorithm is based on knowledge distillation and provides a principled way to maintain the representations of old models while adjusting to new tasks effectively. The proposed method estimates the relationship between the representation changes and the resulting loss increases incurred by model updates. It minimizes the upper bound of the loss increases using the representations, which exploits the estimated importance of each feature map within a backbone model. Based on the importance, the model restricts updates of importa

nt features for robustness while allowing changes in less critical features for flexibility. This optimization strategy effectively alleviates the notorious catastrophic forgetting problem despite the limited accessibility of data in the previous tasks. The experimental results show significant accuracy improvement of the proposed algorithm over the existing methods on the standard datasets. Code is available

****

Learning Program Representations for Food Images and Cooking Recipes

Dim P. Papadopoulos, Enrique Mora, Nadiia Chepurko, Kuan Wei Huang, Ferda Ofli, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16559-16569

In this paper, we are interested in modeling a how-to instructional procedure, such as a cooking recipe, with a meaningful and rich high-level representation. Specifically, we propose to represent cooking recipes and food images as cooking programs. Programs provide a structured representation of the task, capturing cooking semantics and sequential relationships of actions in the form of a graph. This allows them to be easily manipulated by users and executed by agents. To this end, we build a model that is trained to learn a joint embedding between recipes and food images via self-supervision and jointly generate a program from this embedding as a sequence. To validate our idea, we crowdsource programs for cooking recipes and show that: (a) projecting the image-recipe embeddings into programs leads to better cross-modal retrieval results; (b) generating programs from images leads to better recognition results compared to predicting raw cooking instructions; and (c) we can generate food images by manipulating programs via optimizing the latent code of a GAN. Code, data, and models are available online.

****

Bending Graphs: Hierarchical Shape Matching Using Gated Optimal Transport

Mahdi Saleh, Shun-Cheng Wu, Luca Cosmo, Nassir Navab, Benjamin Busam, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11757-11767

Shape matching has been a long-studied problem for the computer graphics and vision community. The objective is to predict a dense correspondence between meshes that have a certain degree of deformation. Existing methods either consider the local description of sampled points or discover correspondences based on global shape information. In this work, we investigate a hierarchical learning design, to which we incorporate local patch-level information and global shape-level structures. This flexible representation enables correspondence prediction and provides rich features for the matching stage. Finally, we propose a novel optimal transport solver by recurrently updating features on non-confident nodes to learn globally consistent correspondences between the shapes. Our results on publicly available datasets suggest robust performance in presence of severe deformations without the need of extensive training or refinement.

****

Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, Prem Natarajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5067-5077

Outside-knowledge visual question answering (OK-VQA) requires the agent to comprehend the image, make use of relevant knowledge from the entire web, and digest all the information to answer the question. Most previous works address the problem by first fusing the image and question in the multi-modal space, which is inflexible for further fusion with a vast amount of external knowledge. In this paper, we call for an alternative paradigm for the OK-VQA task, which transforms the image into plain text, so that we can enable knowledge passage retrieval, and generative question-answering in the natural language space. This paradigm takes advantage of the sheer volume of gigantic knowledge bases and the richness of pre-trained language models. A Transform-Retrieve-Generate framework (TRiG) framework is proposed, which can be plug-and-played with alternative image-to-text models and textual knowledge bases. Experimental results show that our TRiG frame

work outperforms all state-of-the-art supervised methods by at least 11.1% absol
ute margin.
****

Federated Learning With Position-Aware Neurons
Xin-Chun Li, Yi-Chu Xu, Shaoming Song, Bingshuai Li, Yinchuan Li, Yunfeng Shao,
De-Chuan Zhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2022, pp. 10082-10091
Federated Learning (FL) fuses collaborative models from local nodes without cent
ralizing users' data. The permutation invariance property of neural networks and
 the non-i.i.d. data across clients make the locally updated parameters imprecis
ely aligned, disabling the coordinate-based parameter averaging. Traditional neu
rons do not explicitly consider position information. Hence, we propose Position
-Aware Neurons (PANs) as an alternative, fusing position-related values (i.e., p
osition encodings) into neuron outputs. PANs couple themselves to their position
s and minimize the possibility of dislocation, even updating on heterogeneous da
ta. We turn on/off PANs to disable/enable the permutation invariance property of
 neural networks. PANs are tightly coupled with positions when applied to FL, ma
king parameters across clients pre-aligned and facilitating coordinate-based par
ameter averaging. PANs are algorithm-agnostic and could universally improve exis
ting FL algorithms. Furthermore, "FL with PANs" is simple to implement and compu
tationally friendly.
****

Fair Contrastive Learning for Facial Attribute Classification
Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, Hyeran Byun; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2022, pp. 10389-10398
Learning visual representation of high quality is essential for image classifica
tion. Recently, a series of contrastive representation learning methods have ach
ieved preeminent success. Particularly, SupCon outperformed the dominant methods
 based on cross-entropy loss in representation learning. However, we notice that
 there could be potential ethical risks in supervised contrastive learning. In t
his paper, we for the first time analyze unfairness caused by supervised contras
tive learning and propose a new Fair Supervised Contrastive Loss (FSCL) for fair
 visual representation learning. Inheriting the philosophy of supervised contras
tive learning, it encourages representation of the same class to be closer to ea
ch other than that of different classes, while ensuring fairness by penalizing t
he inclusion of sensitive attribute information in representation. In addition,
we introduce a group-wise normalization to diminish the disparities of intra-gro
up compactness and inter-class separability between demographic groups that arou
se unfair classification. Through extensive experiments on CelebA and UTK Face,
we validate that the proposed method significantly outperforms SupCon and existi
ng state-of-the-art methods in terms of the trade-off between top-1 accuracy and
 fairness. Moreover, our method is robust to the intensity of data bias and effe
ctively works in incomplete supervised settings. Our code is available at https:
//github.com/sungho-CoolG/FSCL
****

MDAN: Multi-Level Dependent Attention Network for Visual Emotion Analysis
Liwen Xu, Zhengtao Wang, Bin Wu, Simon Lui; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9479-9488
Visual Emotion Analysis (VEA) is attracting increasing attention. One of the big
gest challenges of VEA is to bridge the affective gap between visual clues in a
picture and the emotion expressed by the picture. As the granularity of emotions
 increases, the affective gap increases as well. Existing deep approaches try to
 bridge the gap by directly learning discrimination among emotions globally in o
ne shot without considering the hierarchical relationship among emotions at diff
erent affective levels and the affective level of emotions to be classified. In
this paper, we present the Multi-level Dependent Attention Network (MDAN) with t
wo branches, to leverage the emotion hierarchy and the correlation between diffe
rent affective levels and semantic levels. The bottom-up branch directly learns
emotions at the highest affective level and strictly follows the emotion hierarc

hy while predicting emotions at lower affective levels. In contrast, the top-dow n branch attempt to disentangle the affective gap by one-to-one mapping between semantic levels and affective levels, namely, Affective Semantic Mapping. At eac h semantic level, a local classifier learns discrimination among emotions at the corresponding affective level. Finally, We integrate global learning and local learning into a unified deep framework and optimize the network simultaneously. Moreover, to properly extract and leverage channel dependencies and spatial atte ntion while disentangling the affective gap, we carefully designed two attention modules: the Multi-head Cross Channel Attention module and the Level-dependent Class Activation Map module. Finally, the proposed deep framework obtains new st ate-of-the-art performance on six VEA benchmarks, where it outperforms existing state-of-the-art methods by a large margin, e.g., +3.85% on the WEBEmo dataset a t 25 classes classification accuracy.

****

Nested Hyperbolic Spaces for Dimensionality Reduction and Hyperbolic NN Design

Xiran Fan, Chun-Hao Yang, Baba C. Vemuri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 356-365

Hyperbolic neural networks have been popular in the recent past due to their abi lity to represent hierarchical data sets effectively and efficiently. The challe nge in developing these networks lies in the nonlinearity of the embedding space namely, the Hyperbolic space. Hyperbolic space is a homogeneous Riemannian mani fold of the Lorentz group which is a semi-Riemannian manifold, i.e. a manifold e quipped with an indefinite metric. Most existing methods (with some exceptions) use local linearization to define a variety of operations paralleling those used in traditional deep neural networks in Euclidean spaces. In this paper, we pres ent a novel fully hyperbolic neural network which uses the concept of projection s (embeddings) followed by an intrinsic aggregation and a nonlinearity all withi n the hyperbolic space. The novelty here lies in the projection which is designe d to project data on to a lower-dimensional embedded hyperbolic space and hence leads to a nested hyperbolic space representation independently useful for dimen sionality reduction. The main theoretical contribution is that the proposed embe dding is proved to be isometric and equivariant under the Lorentz transformation s, which are the natural isometric transformations in hyperbolic spaces. This pr ojection is computationally efficient since it can be expressed by simple linear operations, and, due to the aforementioned equivariance property, it allows for weight sharing. The nested hyperbolic space representation is the core componen t of our network and therefore, we first compare this representation - independe nt of the network - with other dimensionality reduction methods such as tangent PCA, principal geodesic analysis (PGA) and HoroPCA. Based on this equivariant em bedding, we develop a novel fully hyperbolic graph convolutional neural network architecture to learn the parameters of the projection. Finally, we present expe riments demonstrating comparative performance of our network on several publicly available data sets.

****

BNUDC: A Two-Branched Deep Neural Network for Restoring Images From Under-Displa y Cameras

Jaihyun Koh, Jangho Lee, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1950-1959

The images captured by under-display cameras (UDCs) are degraded by the screen i n front of them. We model this degradation in terms of a) diffraction by the pix el grid, which attenuates high-spatial-frequency components of the image; and b) diffuse intensity and color changes caused by the multiple thin-film layers in an OLED, which modulate the low-spatial-frequency components of the image. We in troduce a deep neural network with two branches to reverse each type of degradat ion, which is more effective than performing both restorations in a single forwa rd network. We also propose an affine transform connection to replace the skip c onnection used in most existing DNNs for restoring UDC images. Confining the sol ution space to the linear transform domain reduces the blurring caused by convol ution; and any gross color shift in the training images is eliminated by inverse color filtering. Trained on three datasets of UDC images, our network outperfor

med existing methods in terms of measures of distortion and of perceived image q
uality.
****

RGB-Depth Fusion GAN for Indoor Depth Completion
Haowen Wang, Mingyuan Wang, Zhengping Che, Zhiyuan Xu, Xiuquan Qiao, Mengshi Qi,
 Feifei Feng, Jian Tang; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2022, pp. 6209-6218
The raw depth image captured by the indoor depth sensor usually has an extensive
 range of missing depth values due to inherent limitations such as the inability
 to perceive transparent objects and limited distance range. The incomplete dept
h map burdens many downstream vision tasks, and a rising number of depth complet
ion methods have been proposed to alleviate this issue. While most existing meth
ods can generate accurate dense depth maps from sparse and uniformly sampled dep
th maps, they are not suitable for complementing the large contiguous regions of
 missing depth values, which is common and critical. In this paper, we design a
novel two-branch end-to-end fusion network, which takes a pair of RGB and incomp
leted depth images as input to predict a dense and completed depth map. The firs
t branch employs an encoder-decoder structure to regress the local dense depth v
alues from the raw depth map, with the help of local guidance information extrac
ted from the RGB image. In the other branch, we propose an RGB-depth fusion GAN
to transfer the RGB image to the fine-grained textured depth map. We adopt adapt
ive fusion modules named W-AdaIN to propagate the features across the two branch
es, and we append a confidence fusion head to fuse the two outputs of the branch
es for the final depth map. Extensive experiments on NYU-Depth V2 and SUN RGB-D
demonstrate that our proposed method clearly improves the depth completion perfo
rmance, especially in a more realistic setting of indoor environments with the h
elp of the pseudo depth map.
****

Training Object Detectors From Scratch: An Empirical Study in the Era of Vision
Transformer
Weixiang Hong, Jiangwei Lao, Wang Ren, Jian Wang, Jingdong Chen, Wei Chu; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2022, pp. 4662-4671
Modeling in computer vision has long been dominated by convolutional neural netw
orks (CNNs). Recently, in light of the excellent performances of self-attention
mechanism in the language field, transformers tailored for visual data have draw
n numerous attention and triumphed CNNs in various vision tasks. These vision tr
ansformers heavily rely on large-scale pre-training to achieve competitive accur
acy, which not only hinders the freedom of architectural design in downstream ta
sks like object detection, but also causes learning bias and domain mismatch in
the fine-tuning stages. To this end, we aim to get rid of the "pre-train & fine-
tune" paradigm of vision transformer and train transformer based object detector
 from scratch. Some earlier work in the CNNs era have successfully trained CNNs
based detectors without pre-training, unfortunately, their findings do not gener
alize well when the backbone is switched from CNNs to vision transformer. Instea
d of proposing a specific vision transformer based detector, in this work, our g
oal is to reveal the insights of training vision transformer based detectors fro
m scratch. In particular, we expect those insights can help other researchers an
d practitioners, and inspire more interesting research in other fields, such as
semantic segmentation, visual-linguistic pre-training, etc. One of the key findi
ngs is that both architectural changes and more epochs play critical roles in tr
aining vision transformer based detectors from scratch. Experiments on MS COCO d
atasets demonstrate that vision transformer based detectors trained from scratch
 can also achieve similar performances to their counterparts with ImageNet pre-t
raining.
****

RCL: Recurrent Continuous Localization for Temporal Action Detection
Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13566-13575
Temporal representation is the cornerstone of modern action detection techniques

. State-of-the-art methods mostly rely on a dense anchoring scheme, where anchors are sampled uniformly over the temporal domain with a discretized grid, and then regress the accurate boundaries. In this paper, we revisit this foundational stage and introduce Recurrent Continuous Localization (RCL), which learns a fully continuous anchoring representation. Specifically, the proposed representation builds upon an explicit model conditioned with video embeddings and temporal coordinates, which ensure the capability of detecting segments with arbitrary length. To optimize the continuous representation, we develop an effective scale-invariant sampling strategy and recurrently refine the prediction in subsequent iterations. Our continuous anchoring scheme is fully differentiable, allowing to be seamlessly integrated into existing detectors, e.g., BMN and G-TAD. Extensive experiments on two benchmarks demonstrate that our continuous representation steadily surpasses other discretized counterparts by 2% mAP. As a result, RCL achieves 52.9% mAP@0.5 on THUMOS14 and 37.65% mAP on ActivtiyNet v1.3, outperforming all existing single-model detectors.

****

C2SLR: Consistency-Enhanced Continuous Sign Language Recognition

Ronglai Zuo, Brian Mak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5131-5140

The backbone of most deep-learning-based continuous sign language recognition (CSLR) models consists of a visual module, a sequential module, and an alignment module. However, such CSLR backbones are hard to be trained sufficiently with a single connectionist temporal classification loss. In this work, we propose two auxiliary constraints to enhance the CSLR backbones from the perspective of consistency. The first constraint aims to enhance the visual module, which easily suffers from the insufficient training problem. Specifically, since sign languages convey information mainly with signers' faces and hands, we insert a keypoint-guided spatial attention module into the visual module to enforce it to focus on informative regions, i.e., spatial attention consistency. Nevertheless, only enhancing the visual module may not fully exploit the power of the backbone. Motivated by that both the output features of the visual and sequential modules represent the same sentence, we further impose a sentence embedding consistency constraint between them to enhance the representation power of both the features. Experimental results over three representative backbones validate the effectiveness of the two constraints. More remarkably, with a transformer-based backbone, our model achieves state-of-the-art or competitive performance on three benchmarks, PHOENIX-2014, PHOENIX-2014-T, and CSL.

****

Human Trajectory Prediction With Momentary Observation

Jianhua Sun, Yuxuan Li, Liang Chai, Hao-Shu Fang, Yong-Lu Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6467-6476

Human trajectory prediction task aims to analyze human future movements given their past status, which is a crucial step for many autonomous systems such as self-driving cars and social robots. In real-world scenarios, it is unlikely to obtain sufficiently long observations at all times for prediction, considering inevitable factors such as tracking losses and sudden events. However, the problem of trajectory prediction with limited observations has not drawn much attention in previous work. In this paper, we study a task named momentary trajectory prediction, which reduces the observed history from a long time sequence to an extreme situation of two frames, one frame for social and scene contexts and both frames for the velocity of agents. We perform a rigorous study of existing state-of-the-art approaches in this challenging setting on two widely used benchmarks. We further propose a unified feature extractor, along with a novel pre-training mechanism, to capture effective information within the momentary observation. Our extractor can be adopted in existing prediction models and substantially boost their performance of momentary trajectory prediction. We hope our work will pave the way for more responsive, precise and robust prediction approaches, an important step toward real-world autonomous systems.

****

FoggyStereo: Stereo Matching With Fog Volume Representation
Chengtang Yao, Lidong Yu; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2022, pp. 13043-13052
Stereo matching in foggy scenes is challenging as the scattering effect of fog b
lurs the image and makes the matching ambiguous. Prior methods deem the fog as n
oise and discard it before matching. Different from them, we propose to explore
depth hints from fog and improve stereo matching via these hints. The exploratio
n of depth hints is designed from the perspective of rendering. The rendering is
conducted by reversing the atmospheric scattering process and removing the fog
within a selected depth range. The quality of the rendered image reflects the co
rrectness of the selected depth, as the closer it is to the real depth, the clea
rer the rendered image is. We introduce a fog volume representation to collect t
hese depth hints from the fog. We construct the fog volume by stacking images re
ndered with depths computed from disparity candidates that are also used to buil
d the cost volume. We fuse the fog volume with cost volume to rectify the ambigu
ous matching caused by fog. Experiments show that our fog volume representation
significantly promotes the SOTA result on foggy scenes by 10% ~ 30% while mainta
ining a comparable performance in clear scenes.
****
Trajectory Optimization for Physics-Based Reconstruction of 3D Human Pose From M
onocular Video
Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, Cristian Sminchisescu; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 13106-13115
We focus on the task of estimating a physically plausible articulated human moti
on from monocular video. Existing approaches that do not consider physics often
produce temporally inconsistent output with motion artifacts, while state-of-the
-art physics-based approaches have either been shown to work only in controlled
laboratory conditions or consider simplified body-ground contact limited to feet
. This paper explores how these shortcomings can be addressed by directly incorp
orating a fully-featured physics engine into the pose estimation process. Given
an uncontrolled, real-world scene as input, our approach estimates the ground-pl
ane location and the dimensions of the physical body model. It then recovers the
physical motion by performing trajectory optimization. The advantage of our for
mulation is that it readily generalizes to a variety of scenes that might have d
iverse ground properties and supports any form of self-contact and contact betwe
en the articulated body and scene geometry. We show that our approach achieves c
ompetitive results with respect to existing physics-based methods on the Human3.
6M benchmark, while being directly applicable without re-training to more comple
x dynamic motions from the AIST benchmark and to uncontrolled internet videos.
****
Directional Self-Supervised Learning for Heavy Image Augmentations
Yalong Bai, Yifan Yang, Wei Zhang, Tao Mei; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16692-16701
Despite the large augmentation family, only a few cherry-picked robust augmentat
ion policies are beneficial to self-supervised image representation learning. In
this paper, we propose a directional self-supervised learning paradigm (DSSL),
which is compatible with significantly more augmentations. Specifically, we adap
t heavy augmentation policies after the views lightly augmented by standard augm
entations, to generate harder view (HV). HV usually has a higher deviation from
the original image than the lightly augmented standard view (SV). Unlike previou
s methods equally pairing all augmented views to symmetrically maximize their si
milarities, DSSL treats augmented views of the same instance as a partially orde
red set (with directions as SV\leftrightarrow SV, SV\leftarrowHV), and then equi
ps a directional objective function respecting to the derived relationships amon
g views. DSSL can be easily implemented with a few lines of codes and is highly
flexible to popular self-supervised learning frameworks, including SimCLR, SimSi
am, BYOL. Extensive experimental results on CIFAR and ImageNet demonstrated that
DSSL can stably improve various baselines with compatibility to a wider range o
f augmentations. Code is available at: https://github.com/Yif-Yang/DSSL.

****
Lifelong Unsupervised Domain Adaptive Person Re-Identification With Coordinated Anti-Forgetting and Adaptation
Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14288-14297
Unsupervised domain adaptive person re-identification (ReID) has been extensively investigated to mitigate the adverse effects of domain gaps. Those works assume the target domain data can be accessible all at once. However, for the real-world streaming data, this hinders the timely adaptation to changing data statistics and sufficient exploitation of increasing samples. In this paper, to address more practical scenarios, we propose a new task, Lifelong Unsupervised Domain Adaptive (LUDA) person ReID. This is challenging because it requires the model to continuously adapt to unlabeled data in the target environments while alleviating catastrophic forgetting for such a fine-grained person retrieval task. We design an effective scheme for this task, dubbed CLUDA-ReID, where the anti-forgetting is harmoniously coordinated with the adaptation. Specifically, a meta-based Coordinated Data Replay strategy is proposed to replay old data and update the network with a coordinated optimization direction for both adaptation and memorization. Moreover, we propose Relational Consistency Learning for old knowledge distillation/inheritance in line with the objective of retrieval-based tasks. We set up two evaluation settings to simulate the practical application scenarios. Extensive experiments demonstrate the effectiveness of our CLUDA-ReID for both scenarios with stationary target streams and scenarios with dynamic target streams.
****
No-Reference Point Cloud Quality Assessment via Domain Adaptation
Qi Yang, Yipeng Liu, Siheng Chen, Yiling Xu, Jun Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21179-21188
We present a novel no-reference quality assessment metric, the image transferred point cloud quality assessment (IT-PCQA), for 3D point clouds. For quality assessment, deep neural network (DNN) has shown compelling performance on no-reference metric design. However, the most challenging issue for no-reference PCQA is that we lack large-scale subjective databases to drive robust networks. Our motivation is that the human visual system (HVS) is the decision-maker regardless of the type of media for quality assessment. Leveraging the rich subjective scores of the natural images, we can quest the evaluation criteria of human perception via DNN and transfer the capability of prediction to 3D point clouds. In particular, we treat natural images as the source domain and point clouds as the target domain, and infer point cloud quality via unsupervised adversarial domain adaptation. To extract effective latent features and minimize the domain discrepancy, we propose a hierarchical feature encoder and a conditional-discriminative network. Considering that the ultimate purpose is regressing objective score, we introduce a novel conditional cross entropy loss in the conditional-discriminative network to penalize the negative samples which hinder the convergence of the quality regression network. Experimental results show that the proposed method can achieve higher performance than traditional no-reference metrics, even comparable results with full-reference metrics. The proposed method also suggests the feasibility of assessing the quality of specific media content without the expensive and cumbersome subjective evaluations. Code is available at https://github.com/Qi-Yangsjtu/IT-PCQA.
****
Generating Representative Samples for Few-Shot Classification
Jingyi Xu, Hieu Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9003-9013
Few-shot learning (FSL) aims to learn new categories with a few visual samples per class. Few-shot class representations are often biased due to data scarcity. To mitigate this issue, we propose to generate visual samples based on semantic embeddings using a conditional variational autoencoder (CVAE) model. We train this CVAE model on base classes and use it to generate features for novel classes.

More importantly, we guide this VAE to strictly generate representative samples by removing non-representative samples from the base training set when training the CVAE model. We show that this training scheme enhances the representativeness of the generated samples and therefore, improves the few-shot classification results. Experimental results show that our method improves three FSL baseline methods by substantial margins, achieving state-of-the-art few-shot classification performance on miniImageNet and tieredImageNet datasets for both 1-shot and 5-shot settings.
****

Comprehending and Ordering Semantics for Image Captioning
Yehao Li, Yingwei Pan, Ting Yao, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17990-17999
Comprehending the rich semantics in an image and ordering them in linguistic order are essential to compose a visually-grounded and linguistically coherent description for image captioning. Modern techniques commonly capitalize on a pre-trained object detector/classifier to mine the semantics in an image, while leaving the inherent linguistic ordering of semantics under-exploited. In this paper, we propose a new recipe of Transformer-style structure, namely Comprehending and Ordering Semantics Networks (COS-Net), that novelly unifies an enriched semantic comprehending and a learnable semantic ordering processes into a single architecture. Technically, we initially utilize a cross-modal retrieval model to search the relevant sentences of each image, and all words in the searched sentences are taken as primary semantic cues. Next, a novel semantic comprehender is devised to filter out the irrelevant semantic words in primary semantic cues, and meanwhile infer the missing relevant semantic words visually grounded in the image. After that, we feed all the screened and enriched semantic words into a semantic ranker, which learns to allocate all semantic words in linguistic order as humans. Such sequence of ordered semantic words are further integrated with visual tokens of images to trigger sentence generation. Empirical evidences show that COS-Net clearly surpasses the state-of-the-art approaches on COCO and achieves to-date the best CIDEr score of 141.1% on Karpathy test split. Source code is available at https://github.com/YehLi/xmodaler/tree/master/configs/image_caption/cosnet.
****

Dynamic Scene Graph Generation via Anticipatory Pre-Training
Yiming Li, Xiaoshan Yang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13874-13883
Humans can not only see the collection of objects in visual scenes, but also identify the relationship between objects. The visual relationship in the scene can be abstracted into the semantic representation of triple <subject, predicate, object> and thus results in a scene graph, which can convey a lot of information for visual understanding. Due to the motion of objects, the visual relationship between two objects in videos may vary, which makes the task of dynamically generating scene graphs from videos more complicated and challenging than the conventional image-based static scene graph generation. Inspired by the ability of humans to infer the visual relationship, we propose a novel anticipatory pre-training paradigm based on Transformer to explicitly model the temporal correlation of visual relationships in different frames to improve dynamic scene graph generation. In pre-training stage, the model predicts the visual relationships of current frame based on the previous frames by extracting intra-frame spatial information with a spatial encoder and inter-frame temporal correlations with a temporal encoder. In the fine-tuning stage, we reuse the spatial encoder and the temporal decoder and combine the information of the current frame to predict the visual relationship. Extensive experiments demonstrate that our method achieves state-of-the-art performance on Action Genome dataset.
****

A Large-Scale Comprehensive Dataset and Copy-Overlap Aware Evaluation Protocol for Segment-Level Video Copy Detection
Sifeng He, Xudong Yang, Chen Jiang, Gang Liang, Wei Zhang, Tan Pan, Qing Wang, Furong Xu, Chunguang Li, JinXiong Liu, Hui Xu, Kaiming Huang, Yuan Cheng, Feng Qi

an, Xiaobo Zhang, Lei Yang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2022, pp. 21086-21095
In this paper, we introduce VCSL (Video Copy Segment Localization), a new compre
hensive segment-level annotated video copy dataset. Compared with existing copy
detection datasets restricted by either video-level annotation or small-scale, V
CSL not only has two orders of magnitude more segment-level labelled data, with
160k realistic video copy pairs containing more than 280k localized copied segme
nt pairs, but also covers a variety of video categories and a wide range of vide
o duration. All the copied segments inside each collected video pair are manuall
y extracted and accompanied by precisely annotated starting and ending timestamp
s. Alongside the dataset, we also propose a novel evaluation protocol that bette
r measures the prediction accuracy of copy overlapping segments between a video
pair and shows improved adaptability in different scenarios. By benchmarking sev
eral baseline and state-of-the-art segment-level video copy detection methods wi
th the proposed dataset and evaluation metric, we provide a comprehensive analys
is that uncovers the strengths and weaknesses of current approaches, hoping to o
pen up promising directions for future works. The VCSL dataset, metric and bench
mark codes are all publicly available at https://github.com/alipay/VCSL.
****

GaTector: A Unified Framework for Gaze Object Prediction
Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, Zhijie Zhang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 19588-19597
Gaze object prediction is a newly proposed task that aims to discover the object
s being stared at by humans. It is of great application significance but still l
acks a unified solution framework. An intuitive solution is to incorporate an ob
ject detection branch into an existing gaze prediction method. However, previous
 gaze prediction methods usually use two different networks to extract features
from scene image and head image, which would lead to heavy network architecture
and prevent each branch from joint optimization. In this paper, we build a novel
 framework named GaTector to tackle the gaze object prediction problem in a unif
ied way. Particularly, a specific-general-specific (SGS) feature extractor is fi
rstly proposed to utilize a shared backbone to extract general features for both
 scene and head images. To better consider the specificity of inputs and tasks,
SGS introduces two input-specific blocks before the shared backbone and three ta
sk-specific blocks after the shared backbone. Specifically, a novel Defocus laye
r is designed to generate object-specific features for the object detection task
 without losing information or requiring extra computations. Moreover, the energ
y aggregation loss is introduced to guide the gaze heatmap to concentrate on the
 stared box. In the end, we propose a novel wUoC metric that can reveal the diff
erence between boxes even when they share no overlapping area. Extensive experim
ents on the GOO dataset verify the superiority of our method in all three tracks
, i.e. object detection, gaze estimation, and gaze object prediction.
****

ELIC: Efficient Learned Image Compression With Unevenly Grouped Space-Channel Co
ntextual Adaptive Coding
Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, Yan Wang; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 5718-5727
Recently, learned image compression techniques have achieved remarkable performa
nce, even surpassing the best manually designed lossy image coders. They are pro
mising to be large-scale adopted. For the sake of practicality, a thorough inves
tigation of the architecture design of learned image compression, regarding both
 compression performance and running speed, is essential. In this paper, we firs
t propose uneven channel-conditional adaptive coding, motivated by the observati
on of energy compaction in learned image compression. Combining the proposed une
ven grouping model with existing context models, we obtain a spatial-channel con
textual adaptive model to improve the coding performance without damage to runni
ng speed. Then we study the structure of the main transform and propose an effic
ient model, ELIC, to achieve state-of-the-art speed and compression ability. Wit

h superior performance, the proposed model also supports extremely fast preview decoding and progressive decoding, which makes the coming application of learning-based image compression more promising.
****

CSWin Transformer: A General Vision Transformer Backbone With Cross-Shaped Windows
Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12124-12134
We present CSWin Transformer, an efficient and effective Transformer-based backbone for general-purpose vision tasks. A challenging issue in Transformer design is that global self-attention is very expensive to compute whereas local self-attention often limits the field of interactions of each token. To address this issue, we develop the Cross-Shaped Window self-attention mechanism for computing self-attention in the horizontal and vertical stripes in parallel that form a cross-shaped window, with each stripe obtained by splitting the input feature into stripes of equal width. We provide a mathematical analysis of the effect of the stripe width and vary the stripe width for different layers of the Transformer network which achieves strong modeling capability while limiting the computation cost. We also introduce Locally-enhanced Positional Encoding (LePE), which handles the local positional information better than existing encoding schemes. LePE naturally supports arbitrary input resolutions and is thus especially effective and friendly for downstream tasks. Incorporated with these designs and a hierarchical structure, CSWin Transformer demonstrates competitive performance on common vision tasks. Specifically, it achieves 85.4% Top-1 accuracy on ImageNet-1K without any extra training data or label, 53.9 box AP and 46.4 mask AP on the COCO detection task, and 51.7 mIOU on the ADE20K semantic segmentation task, surpassing previous state-of-the-art Swin Transformer backbone by +1.2, +2.0, +1.4, and +2.0 respectively under the similar FLOPs setting. By further pretraining on the larger dataset ImageNet-21K, we achieve 87.5% Top-1 accuracy on ImageNet-1K and state-of-the-art segmentation performance on ADE20K with 55.7 mIoU.
****

LaTr: Layout-Aware Transformer for Scene-Text VQA
Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, R. Manmatha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16548-16558
We propose a novel multimodal architecture for Scene Text Visual Question Answering (STVQA), named Layout-Aware Transformer (LaTr). The task of STVQA requires models to reason over different modalities. Thus, we first investigate the impact of each modality, and reveal the importance of the language module, especially when enriched with layout information. Accounting for this, we propose a single objective pre-training scheme that requires only text and spatial cues. We show that applying this pre-training scheme on scanned documents has certain advantages over using natural images, despite the domain gap. Scanned documents are easy to procure, text-dense and have a variety of layouts, helping the model learn various spatial cues (e.g. left-of, below etc.) by tying together language and layout information. Compared to existing approaches, our method performs vocabulary-free decoding and, as shown, generalizes well beyond the training vocabulary. We further demonstrate that LaTr improves robustness towards OCR errors, a common reason for failure cases in STVQA. In addition, by leveraging a vision transformer, we eliminate the need for an external object detector. LaTr outperforms state-of-the-art STVQA methods on multiple datasets. In particular, +7.6% on TextVQA, +10.8% on ST-VQA and +4.0% on OCR-VQA (all absolute accuracy numbers).
****

Label Relation Graphs Enhanced Hierarchical Residual Network for Hierarchical Multi-Granularity Classification
Jingzhou Chen, Peng Wang, Jian Liu, Yuntao Qian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4858-4867
Hierarchical multi-granularity classification (HMC) assigns hierarchical multi-granularity labels to each object and focuses on encoding the label hierarchy, e.

g., ["Albatross", "Laysan Albatross"] from coarse-to-fine levels. However, the definition of what is fine-grained is subjective, and the image quality may affect the identification. Thus, samples could be observed at any level of the hierarchy, e.g., ["Albatross"] or ["Albatross", "Laysan Albatross"], and examples discerned at coarse categories are often neglected in the conventional setting of HMC. In this paper, we study the HMC problem in which objects are labeled at any level of the hierarchy. The essential designs of the proposed method are derived from two motivations: (1) learning with objects labeled at various levels should transfer hierarchical knowledge between levels; (2) lower-level classes should inherit attributes related to upper-level superclasses. The proposed combinatorial loss maximizes the marginal probability of the observed ground truth label by aggregating information from related labels defined in the tree hierarchy. If the observed label is at the leaf level, the combinatorial loss further imposes the multi-class cross-entropy loss to increase the weight of fine-grained classification loss. Considering the hierarchical feature interaction, we propose a hierarchical residual network (HRN), in which granularity-specific features from parent levels acting as residual connections are added to features of children levels. Experiments on three commonly used datasets demonstrate the effectiveness of our approach compared to the state-of-the-art HMC approaches. The code will be available at https://github.com/MonsterZhZh/HRN.

****

ITSA: An Information-Theoretic Approach to Automatic Shortcut Avoidance and Domain Generalization in Stereo Matching Networks

WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, David Suter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13022-13032

State-of-the-art stereo matching networks trained only on synthetic data often fail to generalize to more challenging real data domains. In this paper, we attempt to unfold an important factor that hinders the networks from generalizing across domains: through the lens of shortcut learning. We demonstrate that the learning of feature representations in stereo matching networks is heavily influenced by synthetic data artefacts (shortcut attributes). To mitigate this issue, we propose an Information-Theoretic Shortcut Avoidance (ITSA) approach to automatically restrict shortcut-related information from being encoded into the feature representations. As a result, our proposed method learns robust and shortcut-invariant features by minimizing the sensitivity of latent features to input variations. To avoid the prohibitive computational cost of direct input sensitivity optimization, we propose an effective yet feasible algorithm to achieve robustness. We show that using this method, state-of-the-art stereo matching networks that are trained purely on synthetic data can effectively generalize to challenging and previously unseen real data scenarios. Importantly, the proposed method enhances the robustness of the synthetic trained networks to the point that they outperform their fine-tuned counterparts (on real data) for challenging out-of-domain stereo datasets.

****

Enhancing Face Recognition With Self-Supervised 3D Reconstruction

Mingjie He, Jie Zhang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4062-4071

Attributed to both the development of deep networks and abundant data, automatic face recognition (FR) has quickly reached human-level capacity in the past few years. However, the FR problem is not perfectly solved in case of uncontrolled illumination and pose. In this paper, we propose to enhance face recognition with a bypass of self-supervised 3D reconstruction, which enforces the neural backbone to focus on the identity-related depth and albedo information while neglects the identity-irrelevant pose and illumination information. Specifically, inspired by the physical model of image formation, we improve the backbone FR network by introducing a 3D face reconstruction loss with two auxiliary networks. The first one estimates the pose and illumination from the input face image while the second one decodes the canonical depth and albedo from the intermediate feature of the FR backbone network. The whole network is trained in end-to-end manner wit

h both classic face identification loss and the loss of 3D face reconstruction w
ith the physical parameters. In this way, the self-supervised reconstruction act
s as a regularization that enables the recognition network to understand faces i
n 3D view, and the learnt features are forced to encode more information of cano
nical facial depth and albedo, which is more intrinsic and beneficial to face re
cognition. Extensive experimental results on various face recognition benchmarks
 show that, without any cost of extra annotations and computations, our method o
utperforms state-of-the-art ones. Moreover, the learnt representations can also
well generalize to other face-related downstream tasks such as the facial attrib
ute recognition with limited labeled data.
****

HeadNeRF: A Real-Time NeRF-Based Parametric Head Model
Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, Juyong Zhang; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2
0374-20384
In this paper, we propose HeadNeRF, a novel NeRF-based parametric head model tha
t integrates the neural radiance field to the parametric representation of the h
uman head. It can render high fidelity head images in real-time on modern GPUs,
and supports directly controlling the generated images' rendering pose and vario
us semantic attributes. Different from existing related parametric models, we us
e the neural radiance fields as a novel 3D proxy instead of the traditional 3D t
extured mesh, which makes that HeadNeRF is able to generate high fidelity images
. However, the computationally expensive rendering process of the original NeRF
hinders the construction of the parametric NeRF model. To address this issue, we
 adopt the strategy of integrating 2D neural rendering to the rendering process
of NeRF and design novel loss terms. As a result, the rendering speed of HeadNeR
F can be significantly accelerated, and the rendering time of one frame is reduc
ed from 5s to 25ms. The well designed loss terms also improve the rendering accu
racy, and the fine-level details of the human head, such as the gaps between tee
th, wrinkles, and beards, can be represented and synthesized by HeadNeRF. Extens
ive experimental results and several applications demonstrate its effectiveness.
 The trained parametric model is available at https://github.com/CrisHY1995/head
nerf.
****

FvOR: Robust Joint Shape and Pose Optimization for Few-View Object Reconstructio
n
Zhenpei Yang, Zhile Ren, Miguel Angel Bautista, Zaiwei Zhang, Qi Shan, Qixing Hu
ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2022, pp. 2497-2507
Reconstructing an accurate 3D object model from a few image observations remains
 a challenging problem in computer vision. State-of-the-art approaches typically
 assume accurate camera poses as input, which could be difficult to obtain in re
alistic settings. In this paper, we present FvOR, a learning-based object recons
truction method that predicts accurate 3D models given a few images with noisy i
nput poses. The core of our approach is a fast and robust multi-view reconstruct
ion algorithm to jointly refine 3D geometry and camera pose estimation using lea
rnable neural network modules. We provide a thorough benchmark of state-of-the-a
rt approaches for this problem on ShapeNet. Our approach achieves best-in-class
results. It is also two orders of magnitude faster than the recent optimization-
based approach IDR.
****

Reduce Information Loss in Transformers for Pluralistic Image Inpainting
Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengc
hen Liu, Lu Yuan, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2022, pp. 11347-11357
Transformers have achieved great success in pluralistic image inpainting recentl
y. However, we find existing transformer based solutions regard each pixel as a
token, thus suffer from information loss issue from two aspects: 1) They downsam
ple the input image into much lower resolutions for efficiency consideration, in
curring information loss and extra misalignment for the boundaries of masked reg

ions. 2) They quantize 2563 RGB pixels to a small number (such as 512) of quanti zed pixels. The indices of quantized pixels are used as tokens for the inputs an d prediction targets of transformer. Although an extra CNN network is used to up sample and refine the low-resolution results, it is difficult to retrieve the lo st information back. To keep input information as much as possible, we propose a new transformer based framework "PUT". Specifically, to avoid input downsamplin g while maintaining the computation efficiency, we design a patch-based auto-enc oder PVQVAE, where the encoder converts the masked image into non-overlapped pat ch tokens and the decoder recovers the masked regions from the inpainted tokens while keeping the unmasked regions unchanged. To eliminate the information loss caused by quantization, an Un-Quantized Transformer (UQ-Transformer) is applied, which directly takes the features from P-VQVAE encoder as input without quantiz ation and regards the quantized tokens only as prediction targets. Extensive exp eriments show that PUT greatly outperforms state-of-the-art methods on image fid elity, especially for large masked regions and complex large-scale datasets.
****

Replacing Labeled Real-Image Datasets With Auto-Generated Contours
Hirokatsu Kataoka, Ryo Hayamizu, Ryosuke Yamada, Kodai Nakashima, Sora Takashima , Xinyu Zhang, Edgar Josafat Martinez-Noriega, Nakamasa Inoue, Rio Yokota; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 21232-21241
In the present work, we show that the performance of formula-driven supervised l earning (FDSL) can match or even exceed that of ImageNet-21k without the use of real images, human-, and self-supervision during the pre-training of Vision Tran sformers (ViTs). For example, ViT-Base pre-trained on ImageNet-21k shows 81.8% t op-1 accuracy when fine-tuned on ImageNet-1k and FDSL shows 82.7% top-1 accuracy when pre-trained under the same conditions (number of images, hyperparameters, and number of epochs). Images generated by formulas avoid the privacy/copyright issues, labeling cost and errors, and biases that real images suffer from, and t hus have tremendous potential for pre-training general models. To understand the performance of the synthetic images, we tested two hypotheses, namely (i) objec t contours are what matter in FDSL datasets and (ii) increased number of paramet ers to create labels affects performance improvement in FDSL pre-training. To te st the former hypothesis, we constructed a dataset that consisted of simple obje ct contour combinations. We found that this dataset can match the performance of fractals. For the latter hypothesis, we found that increasing the difficulty of the pre-training task generally leads to better fine-tuning accuracy.
****

Cross-Modal Transferable Adversarial Attacks From Images to Videos
Zhipeng Wei, Jingjing Chen, Zuxuan Wu, Yu-Gang Jiang; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15064-15073
Recent studies have shown that adversarial examples hand-crafted on one white bo x model can be used to attack other black-box models. Such cross-model transfera bility makes it feasible to perform black-box attacks, which has raised security concerns for real-world DNNs applications. Nevertheless, existing works mostly focus on investigating the adversarial transferability across different deep mod els that share the same modality of input data. The cross-modal transferability of adversarial perturbation has never been explored. This paper investigates the transferability of adversarial perturbation across different modalities, i.e., leveraging adversarial perturbation generated on white-box image models to attac k black-box video models. Specifically, motivated by the observation that the lo w-level feature space between images and video frames are similar, we propose a simple yet effective cross-modal attack method, named as Image To Video (I2V) at tack. I2V generates adversarial frames by minimizing the cosine similarity betwe en features of pre-trained image models from adversarial and benign examples, th en combines the generated adversarial frames to perform black-box attacks on vid eo recognition models. Extensive experiments demonstrate that I2V can achieve hi gh attack success rates on different black-box video recognition models. On Kine tics-400 and UCF-101, I2V achieves an average attack success rate of 77.88% and

65.68%, respectively, which sheds light on the feasibility of cross-modal adversarial attacks.
****

Few Could Be Better Than All: Feature Sampling and Grouping for Scene Text Detection
Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4563-4572
Recently, transformer-based methods have achieved promising progresses in object detection, as they can eliminate the post-processes like NMS and enrich the deep representations. However, these methods cannot well cope with scene text due to its extreme variance of scales and aspect ratios. In this paper, we present a simple yet effective transformer-based architecture for scene text detection. Different from previous approaches that learn robust deep representations of scene text in a holistic manner, our method performs scene text detection based on a few representative features, which avoids the disturbance by background and reduces the computational cost. Specifically, we first select a few representative features at all scales that are highly relevant to foreground text. Then, we adopt a transformer for modeling the relationship of the sampled features, which effectively divides them into reasonable groups. As each feature group corresponds to a text instance, its bounding box can be easily obtained without any post-processing operation. Using the basic feature pyramid network for feature extraction, our method consistently achieves state-of-the-art results on several popular datasets for scene text detection.
****

Do Explanations Explain? Model Knows Best
Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, Nassir Navab; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10244-10253
It is a mystery which input features contribute to a neural network's output. Various explanations methods are proposed in the literature to shed light on the problem. One peculiar observation is that these explanations point to different features as being important. The phenomenon raises the question, which explanation to trust? We propose a framework for evaluating the explanations using the neural network model itself. The framework leverages the network to generate input features that impose a particular behavior on the output. Using the generated features, we devise controlled experimental setups to evaluate whether an explanation method conforms to an axiom. Thus we propose an empirical framework for axiomatic evaluation of explanation methods. We evaluate well-known and promising explanation solutions using the proposed framework. The framework provides a toolset to reveal properties and drawbacks within existing and future explanation solutions.
****

WebQA: Multihop and Multimodal QA
Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, Yonatan Bisk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16495-16504
Scaling Visual Question Answering (VQA) to the open-domain and multi-hop nature of web searches, requires fundamental advances in visual representation learning, knowledge aggregation, and language generation. In this work, we introduce WebQA, a challenging new benchmark that proves difficult for large-scale state-of-the-art models which lack language groundable visual representations for novel objects and the ability to reason, yet trivial for humans. WebQA mirrors the way humans use the web: 1) Ask a question, 2) Choose sources to aggregate, and 3) Produce a fluent language response. This is the behavior we should be expecting from IoT devices and digital assistants. Existing work prefers to assume that a model can either reason about knowledge in images or in text. WebQA includes a secondary text-only QA task to ensure improved visual performance does not come at the cost of language understanding. Our challenge for the community is to create unified multimodal reasoning models that answer questions regardless of the sour

ce modality, moving us closer to digital assistants that not only query language knowledge, but also the richer visual online world.

****

Occlusion-Robust Face Alignment Using a Viewpoint-Invariant Hierarchical Network Architecture

Congcong Zhu, Xintong Wan, Shaorong Xie, Xiaoqiang Li, Yinzheng Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11112-11121

The occlusion problem heavily degrades the localization performance of face alignment. Most current solutions for this problem focus on annotating new occlusion data, introducing boundary estimation, and stacking deeper models to improve the robustness of neural networks. However, the performance degradation of models remains under extreme occlusion (average occlusion of over 50%) because of missing a large amount of facial context information. We argue that exploring neural networks to model the facial hierarchies is a more promising method for dealing with extreme occlusion. Surprisingly, in recent studies, little effort has been devoted to representing the facial hierarchies using neural networks. This paper proposes a new network architecture called GlomFace to model the facial hierarchies against various occlusions, which draws inspiration from the viewpoint-invariant hierarchy of facial structure. Specifically, GlomFace is functionally divided into two modules: the part-whole hierarchical module and the whole-part hierarchical module. The former captures the part-whole hierarchical dependencies of facial parts to suppress multi-scale occlusion information, whereas the latter injects structural reasoning into neural networks by building the whole-part hierarchical relations among facial parts. As a result, GlomFace has a clear topological interpretation due to its correspondence to the facial hierarchies. Extensive experimental results indicate that the proposed GlomFace performs comparably to existing state-of-the-art methods, especially in cases of extreme occlusion. Models are available at https://github.com/zhuccly/GlomFace-Face-Alignment.

****

BasicVSR++: Improving Video Super-Resolution With Enhanced Propagation and Alignment

Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5972-5981

A recurrent structure is a popular framework choice for the task of video super-resolution. The state-of-the-art method BasicVSR adopts bidirectional propagation with feature alignment to effectively exploit information from the entire input video. In this study, we redesign BasicVSR by proposing second-order grid propagation and flow-guided deformable alignment. We show that by empowering the recurrent framework with enhanced propagation and alignment, one can exploit spatio temporal information across misaligned video frames more effectively. The new components lead to an improved performance under a similar computational constraint. In particular, our model BasicVSR++ surpasses BasicVSR by a significant 0.82 dB in PSNR with similar number of parameters. BasicVSR++ is generalizable to other video restoration tasks, and obtains three champions and one first runner-up in NTIRE 2021 video restoration challenge.

****

IDR: Self-Supervised Image Denoising via Iterative Data Refinement

Yi Zhang, Dasong Li, Ka Lung Law, Xiaogang Wang, Hongwei Qin, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2098-2107

The lack of large-scale noisy-clean image pairs restricts supervised denoising methods' deployment in actual applications. While existing unsupervised methods are able to learn image denoising without ground-truth clean images, they either show poor performance or work under impractical settings (e.g., paired noisy images). In this paper, we present a practical unsupervised image denoising method to achieve state-of-the-art denoising performance. Our method only requires single noisy images and a noise model, which is easily accessible in practical raw image denoising. It performs two steps iteratively: (1) Constructing a noisier-no

isy dataset with random noise from the noise model; (2) training a model on the noisier-noisy dataset and using the trained model to refine noisy images to obtain the targets used in the next round. We further approximate our full iterative method with a fast algorithm for more efficient training while keeping its original high performance. Experiments on real-world, synthetic, and correlated noise show that our proposed unsupervised denoising approach has superior performances over existing unsupervised methods and competitive performance with supervised methods. In addition, we argue that existing denoising datasets are of low quality and contain only a small number of scenes. To evaluate raw image denoising performance in real-world applications, we build a high-quality raw image dataset SenseNoise-500 that contains 500 real-life scenes. The dataset can serve as a strong benchmark for better evaluating raw image denoising. Code and dataset will be released at https://github.com/zhangyi-3/IDR

****

MogFace: Towards a Deeper Appreciation on Face Detection

Yang Liu, Fei Wang, Jiankang Deng, Zhipeng Zhou, Baigui Sun, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4093-4102

Benefiting from the pioneering design of generic object detectors, significant achievements have been made in the field of face detection. Typically, the architectures of the backbone, feature pyramid layer, and detection head module within the face detector all assimilate the excellent experience from general object detectors. However, several effective methods, including label assignment and scale-level data augmentation strategy, fail to maintain consistent superiority when applying on the face detector directly. Concretely, the former strategy involves a vast body of hyper-parameters and the latter one suffers from the challenge of scale distribution bias between different detection tasks, which both limit their generalization abilities. Furthermore, in order to provide accurate face bounding boxes for facial down-stream tasks, the face detector imperatively requires the elimination of false alarms. As a result, practical solutions on label assignment, scale-level data augmentation, and reducing false alarms are necessary for advancing face detectors. In this paper, we focus on resolving three aforementioned challenges that exiting methods are difficult to finish off and present a novel face detector, termed MogFace. In our Mogface, three key components, Adaptive Online Incremental Anchor Mining Strategy, Selective Scale Enhancement Strategy and Hierarchical Context-Aware Module, are separately proposed to boost the performance of face detectors. Finally, to the best of our knowledge, our MogFace is the best face detector on the Wider Face leader-board, achieving all champions across different testing scenarios. The code is available at https://github.com/damo-cv/MogFace.

****

GuideFormer: Transformers for Image Guided Depth Completion

Kyeongha Rho, Jinsung Ha, Youngjung Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6250-6259

Depth completion has been widely studied to predict a dense depth image from its sparse measurement and a single color image. However, most state-of-the-art methods rely on static convolutional neural networks (CNNs) which are not flexible enough for capturing the dynamic nature of input contexts. In this paper, we propose GuideFormer, a fully transformer-based architecture for dense depth completion. We first process sparse depth and color guidance images with separate transformer branches to extract hierarchical and complementary token representations. Each branch consists of a stack of self-attention blocks and has key design features to make our model suitable for the task. We also devise an effective token fusion method based on guided-attention mechanism. It explicitly models information flow between the two branches and captures inter-modal dependencies that cannot be obtained from depth or color image alone. These properties allow GuideFormer to enjoy various visual dependencies and recover precise depth values while preserving fine details. We evaluate GuideFormer on the KITTI dataset containing real-world driving scenes and provide extensive ablation studies. Experimental results demonstrate that our approach significantly outperforms the state-of-th

e-art methods.
****

Multi-Label Iterated Learning for Image Classification With Label Ambiguity
Sai Rajeswar, Pau Rodríguez, Soumye Singhal, David Vazquez, Aaron Courville; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4783-4793
Transfer learning from large-scale pre-trained models has become essential for many computer vision tasks. Recent studies have shown that datasets like ImageNet are weakly labeled since images with multiple object classes present are assigned a single label. This ambiguity biases models towards a single prediction, which could result in the suppression of classes that tend to co-occur in the data. Inspired by language emergence literature, we propose multi-label iterated learning (MILe) to incorporate the inductive biases of multi-label learning from single labels using the framework of iterated learning. MILe is a simple yet effective procedure that builds a multi-label description of the image by propagating binary predictions through successive generations of teacher and student networks with a learning bottleneck. Experiments show that our approach exhibits systematic benefits on ImageNet accuracy as well as ReaL F1 score, which indicates that MILe deals better with label ambiguity than the standard training procedure, even when fine-tuning from self-supervised weights. We also show that MILe is effective reducing label noise, achieving state-of-the-art performance on real-world large-scale noisy data such as WebVision. Furthermore, MILe improves performance in class incremental settings such as IIRC and it is robust to distribution shifts.
****

Region-Aware Face Swapping
Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7632-7641
This paper presents a novel Region-Aware Face Swapping (RAFSwap) network to achieve identity-consistent harmonious high-resolution face generation in a local-global manner: 1) Local Facial Region-Aware (FRA) branch augments local identity-relevant features by introducing the Transformer to effectively model misaligned cross-scale semantic interaction. 2) Global Source Feature-Adaptive (SFA) branch further complements global identity-relevant cues for generating identity-consistent swapped faces. Besides, we propose a Face Mask Predictor (FMP) module incorporated with StyleGAN2 to predict identity-relevant soft facial masks in an unsupervised manner that is more practical for generating harmonious high-resolution faces. Abundant experiments qualitatively and quantitatively demonstrate the superiority of our method for generating more identity-consistent high-resolution swapped faces over SOTA methods, e.g., obtaining 96.70 ID retrieval that outperforms SOTA MegaFS by 5.87.
****

Towards Language-Free Training for Text-to-Image Generation
Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, Tong Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17907-17917
One of the major challenges in training text-to-image generation models is the need of a large number of high-quality text-image pairs. While image samples are often easily accessible, the associated text description typically requires careful human captioning, which is particularly time- and cost-consuming. In this paper, we propose the first work to train text-to-image generation models without any text data. It intelligently leverages the well-aligned cross-modal semantic space of the powerful pre-trained CLIP model: the requirement of text-conditioning is alleviated via generating text features from image features. Extensive experiments are conducted to illustrate the effectiveness of the proposed method. We obtain state-of-the-art results in the standard text-to-image generation tasks. Importantly, the proposed language-free model outperforms most existing models trained with full text-image pairs. Furthermore, our method can be applied in fine-tuning pre-trained models, which saves both training time and cost in traini

ng text-to-image generation models. Our pre-trained model obtains competitive results in zero-shot text-to-image generation on MS-COCO dataset, yet with around only 1% of the model size compared to the recently proposed large DALL-E model.
****

Learning Affinity From Attention: End-to-End Weakly-Supervised Semantic Segmentation With Transformers
Lixiang Ru, Yibing Zhan, Baosheng Yu, Bo Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16846-16855
Weakly-supervised semantic segmentation (WSSS) with image-level labels is an important and challenging task. Due to the high training efficiency, end-to-end solutions for WSSS have received increasing attention from the community. However, current methods are mainly based on convolutional neural networks and fail to explore the global information properly, thus usually resulting in incomplete object regions. In this paper, to address the aforementioned problem, we introduce Transformers, which naturally integrate global information, to generate more integral initial pseudo labels for end-to-end WSSS. Motivated by the inherent consistency between the self-attention in Transformers and the semantic affinity, we propose an Affinity from Attention (AFA) module to learn semantic affinity from the multi-head self-attention (MHSA) in Transformers. The learned affinity is then leveraged to refine the initial pseudo labels for segmentation. In addition, to efficiently derive reliable affinity labels for supervising AFA and ensure the local consistency of pseudo labels, we devise a Pixel-Adaptive Refinement module that incorporates low-level image appearance information to refine the pseudo labels. We perform extensive experiments and our method achieves 66.0% and 38.9% mIoU on the PASCAL VOC 2012 and MS COCO 2014 datasets, respectively, significantly outperforming recent end-to-end methods and several multi-stage competitors. Code is available at https://github.com/rulixiang/afa.
****

Pushing the Envelope of Gradient Boosting Forests via Globally-Optimized Oblique Trees
Magzhan Gabidolla, Miguel Á. Carreira-Perpiñán; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 285-294
Ensemble methods based on decision trees, such as Random Forests or boosted forests, have long been established as some of the most powerful, off-the-shelf machine learning models, and have been widely used in computer vision and other areas. In recent years, a specific form of boosting, gradient boosting (GB), has gained prominence. This is partly because of highly optimized implementations such as XGBoost or LightGBM, which incorporate many clever modifications and heuristics. However, one gaping hole remains unexplored in GB: the construction of individual trees. To date, all successful GB versions use axis-aligned trees trained in a suboptimal way via greedy recursive partitioning. We address this gap by using a more powerful type of trees (having hyperplane splits) and an algorithm that can optimize, globally over all the tree parameters, the objective function that GB dictates. We show, in several benchmarks of image and other data types, that GB forests of these stronger, well-optimized trees consistently exceed the test accuracy of axis-aligned forests from XGBoost, LightGBM and other strong baselines. Further, this happens using many fewer trees and sometimes even fewer parameters overall.
****

Physical Simulation Layer for Accurate 3D Modeling
Mariem Mezghanni, Théo Bodrito, Malika Boulkenafed, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13514-13523
We introduce a novel approach for generative 3D modeling that explicitly encourages the physical and thus functional consistency of the generated shapes. To this end, we advocate the use of online physical simulation as part of learning a generative model. Unlike previous related methods, our approach is trained end-to-end with a fully differentiable physical simulator in the training loop. We accomplish this by leveraging recent advances in differentiable programming, and introducing a fully differentiable point-based physical simulation layer, which ac

curately evaluates the shape's stability when subjected to gravity. We then inco
rporate this layer in a signed distance function (SDF) shape decoder. By augment
ing a conventional SDF decoder with our simulation layer, we demonstrate through
 extensive experiments that online physical simulation improves the accuracy, vi
sual plausibility and physical validity of the resulting shapes, while requiring
 no additional data or annotation effort.
****

Deformable Sprites for Unsupervised Video Decomposition
Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, Noah Snavely; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2022, pp. 2657-2666
We describe a method to extract persistent elements of a dynamic scene from an i
nput video. We represent each scene element as a Deformable Sprite consisting of
 three components: 1) a 2D texture image for the entire video, 2) per-frame mask
s for the element, and 3) non-rigid deformations that map the texture image into
 each video frame. The resulting decomposition allows for applications such as c
onsistent video editing. Deformable Sprites are a type of video auto-encoder mod
el that is optimized on individual videos, and does not require training on a la
rge dataset, nor does it rely on pre-trained models. Moreover, our method does n
ot require object masks or other user input, and discovers moving objects of a w
ider variety than previous work. We evaluate our approach on standard video data
sets and show qualitative results on a diverse array of Internet videos.
****

CamLiFlow: Bidirectional Camera-LiDAR Fusion for Joint Optical Flow and Scene Fl
ow Estimation
Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, Lijun Chen; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p
p. 5791-5801
In this paper, we study the problem of jointly estimating the optical flow and s
cene flow from synchronized 2D and 3D data. Previous methods either employ a com
plex pipeline that splits the joint task into independent stages, or fuse 2D and
 3D information in an "early-fusion" or "late-fusion" manner. Such one-size-fits
-all approaches suffer from a dilemma of failing to fully utilize the characteri
stic of each modality or to maximize the inter-modality complementarity. To addr
ess the problem, we propose a novel end-to-end framework, called CamLiFlow. It c
onsists of 2D and 3D branches with multiple bidirectional connections between th
em in specific layers. Different from previous work, we apply a point-based 3D b
ranch to better extract the geometric features and design a symmetric learnable
operator to fuse dense image features and sparse point features. Experiments sho
w that CamLiFlow achieves better performance with fewer parameters. Our method r
anks 1st on the KITTI Scene Flow benchmark, outperforming the previous art with
1/7 parameters. Code is available at https://github.com/MCG-NJU/CamLiFlow.
****

FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in
Videos
Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifen
g Ge, Wenqiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2022, pp. 20922-20931
Current benchmarks for facial expression recognition (FER) mainly focus on stati
c images, while there are limited datasets for FER in videos. It is still ambigu
ous to evaluate whether performances of existing methods remain satisfactory in
real-world application-oriented scenes. For example, the "Happy" expression with
 high intensity in Talk-Show is more discriminating than the same expression wit
h low intensity in Official-Event. To fill this gap, we build a large-scale mult
i-scene dataset, coined as FERV39k. We analyze the important ingredients of cons
tructing such a novel dataset in three aspects: (1) multi-scene hierarchy and ex
pression class, (2) generation of candidate video clips, (3) trusted manual labe
lling process. Based on these guidelines, we select 4 scenarios subdivided into
22 scenes, annotate 86k samples automatically obtained from 4k videos based on t
he well-designed workflow, and finally build 38,935 video clips labeled with 7 c

lassic expressions. Experiment benchmarks on four kinds of baseline frameworks w ere also provided and further analysis on their performance across different sce nes and some challenges for future research were given. Besides, we systematical ly investigate key components of DFER by ablation studies. The baseline framewor k and our project are available on https://github.com/wangyanckxx/FERV39k.
****

Learning To Detect Mobile Objects From LiDAR Scans Without Labels
Yurong You, Katie Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Harihara n, Mark Campbell, Kilian Q. Weinberger; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1130-1140
Current 3D object detectors for autonomous driving are almost entirely trained o n human-annotated data. Although of high quality, the generation of such data is laborious and costly, restricting them to a few specific locations and object t ypes. This paper proposes an alternative approach entirely based on unlabeled da ta, which can be collected cheaply and in abundance almost everywhere on earth. Our approach leverages several simple common sense heuristics to create an initi al set of approximate seed labels. For example, relevant traffic participants ar e generally not persistent across multiple traversals of the same route, do not fly, and are never under ground. We demonstrate that these seed labels are highl y effective to bootstrap a surprisingly accurate detector through repeated self- training without a single human annotated label. Code is available at https://gi thub.com/YurongYou/MODEST.
****

BNV-Fusion: Dense 3D Reconstruction Using Bi-Level Neural Volume Fusion
Kejie Li, Yansong Tang, Victor Adrian Prisacariu, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 6166-6175
Dense 3D reconstruction from a stream of depth images is the key to many mixed r eality and robotic applications. Although methods based on Truncated Signed Dist ance Function (TSDF) Fusion have advanced the field over the years, the TSDF vol ume representation is confronted with striking a balance between the robustness to noisy measurements and maintaining the level of detail. We present Bi-level N eural Volume Fusion (BNV-Fusion), which leverages recent advances in neural impl icit representations and neural rendering for dense 3D reconstruction. In order to incrementally integrate new depth maps into a global neural implicit represen tation, we propose a novel bi-level fusion strategy that considers both efficien cy and reconstruction quality by design. We evaluate the proposed method on mult iple datasets quantitatively and qualitatively, demonstrating a significant impr ovement over existing methods.
****

Probabilistic Representations for Video Contrastive Learning
Jungin Park, Jiyoung Lee, Ig-Jae Kim, Kwanghoon Sohn; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14711- 14721
This paper presents Probabilistic Video Contrastive Learning, a self-supervised representation learning method that bridges contrastive learning with probabilis tic representation. We hypothesize that the clips composing the video have diffe rent distributions in short-term duration, but can represent the complicated and sophisticated video distribution through combination in a common embedding spac e. Thus, the proposed method represents video clips as normal distributions and combines them into a Mixture of Gaussians to model the whole video distribution. By sampling embeddings from the whole video distribution, we can circumvent the careful sampling strategy or transformations to generate augmented views of the clips, unlike previous deterministic methods that have mainly focused on such s ample generation strategies for contrastive learning. We further propose a stoch astic contrastive loss to learn proper video distributions and handle the inhere nt uncertainty from the nature of the raw video. Experimental results verify tha t our probabilistic embedding stands as a state-of-the-art video representation learning for action recognition and video retrieval on the most popular benchmar ks, including UCF101 and HMDB51.

****
EnvEdit: Environment Editing for Vision-and-Language Navigation
Jialu Li, Hao Tan, Mohit Bansal; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2022, pp. 15407-15417
In Vision-and-Language Navigation (VLN), an agent needs to navigate through the environment based on natural language instructions. Due to limited available dat a for agent training and finite diversity in navigation environments, it is chal lenging for the agent to generalize to new, unseen environments. To address this problem, we propose EnvEdit, a data augmentation method that creates new enviro nments by editing existing environments, which are used to train a more generali zable agent. Our augmented environments can differ from the seen environments in three diverse aspects: style, object appearance, and object classes. Training o n these edit-augmented environments prevents the agent from overfitting to exist ing environments and helps generalize better to new, unseen environments. Empiri cally, on both the Room-to-Room and the multi-lingual Room-Across-Room datasets, we show that our proposed EnvEdit method gets significant improvements in all m etrics on both pre-trained and non-pre-trained VLN agents, and achieves the new state-of-the-art on the test leaderboard. We further ensemble the VLN agents aug mented on different edited environments and show that these edit methods are com plementary.
****
Omnivore: A Single Model for Many Visual Modalities
Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin , Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 16102-16112
Prior work has studied different visual modalities in isolation and developed se parate architectures for recognition of images, videos, and 3D data. Instead, in this paper, we propose a single model which excels at classifying images, video s, and single-view 3D data using exactly the same model parameters. Our 'OMNIVOR E' model leverages the flexibility of transformer-based architectures and is tra ined jointly on classification tasks from different modalities. OMNIVORE is simp le to train, uses off-the-shelf standard datasets, and performs at-par or better than modality-specific models of the same size. A single OMNIVORE model obtains 86.0% on ImageNet, 84.1% on Kinetics, and 67.1% on SUN RGB-D. After finetuning, our models outperform prior work on a variety of vision tasks and generalize ac ross modalities. OMNIVORE's shared visual representation naturally enables cross -modal recognition without access to correspondences between modalities. We hope our results motivate researchers to model visual modalities together.
****
Neural Shape Mating: Self-Supervised Object Assembly With Adversarial Shape Prio rs
Yun-Chun Chen, Haoda Li, Dylan Turpin, Alec Jacobson, Animesh Garg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 12724-12733
Learning to autonomously assemble shapes is a crucial skill for many robotic app lications. While the majority of existing part assembly methods focus on correct ly posing semantic parts to recreate a whole object, we interpret assembly more literally: as mating geometric parts together to achieve a snug fit. By focusing on shape alignment rather than semantic cues, we can achieve across category ge neralization and scaling. In this paper, we introduce a novel task, pairwise 3D geometric shape mating, and propose Neural Shape Mating (NSM) to tackle this pro blem. Given point clouds of two object parts of an unknown category, NSM learns to reason about the fit of the two parts and predict a pair of 3D poses that tig htly mate them together. In addition, we couple the training of NSM with an impl icit shape reconstruction task, making NSM more robust to imperfect point cloud observations. To train NSM, we present a self-supervised data collection pipelin e that generates pairwise shape mating data with ground truth by randomly cuttin g an object mesh into two parts, resulting in a dataset that consists of 200K sh ape mating pairs with numerous object meshes and diverse cut types. We train NSM on the collected dataset and compare it with several point cloud registration m

ethods and one part assembly baseline approach. Extensive experimental results a
nd ablation studies under various settings demonstrate the effectiveness of the
proposed algorithm. Additional material is available at: neural-shape-mating.git
hub.io.
****

Reflash Dropout in Image Super-Resolution
Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, Chao Dong; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6002
-6012
Dropout is designed to relieve the overfitting problem in high-level vision task
s but is rarely applied in low-level vision tasks, like image super-resolution (
SR). As a classic regression problem, SR exhibits a different behaviour as high-
level tasks and is sensitive to the dropout operation. However, in this paper, w
e show that appropriate usage of dropout benefits SR networks and improves the g
eneralization ability. Specifically, dropout is better embedded at the end of th
e network and is significantly helpful for the multi-degradation settings. This
discovery breaks our common sense and inspires us to explore its working mechani
sm. We further use two analysis tools -- one is from recent network interpretati
on works, and the other is specially designed for this task. The analysis result
s provide side proofs to our experimental findings and show us a new perspective
 to understand SR networks.
****

WildNet: Learning Domain Generalized Semantic Segmentation From the Wild
Suhyeon Lee, Hongje Seong, Seongwon Lee, Euntai Kim; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9936-99
46
We present a new domain generalized semantic segmentation network named WildNet,
 which learns domain-generalized features by leveraging a variety of contents an
d styles from the wild. In domain generalization, the low generalization ability
 for unseen target domains is clearly due to overfitting to the source domain. T
o address this problem, previous works have focused on generalizing the domain b
y removing or diversifying the styles of the source domain. These alleviated ove
rfitting to the source-style but overlooked overfitting to the source-content. I
n this paper, we propose to diversify both the content and style of the source d
omain with the help of the wild. Our main idea is for networks to naturally lear
n domain-generalized semantic information from the wild. To this end, we diversi
fy styles by augmenting source features to resemble wild styles and enable netwo
rks to adapt to a variety of styles. Furthermore, we encourage networks to learn
 class-discriminant features by providing semantic variations borrowed from the
wild to source contents in the feature space. Finally, we regularize networks to
 capture consistent semantic information even when both the content and style of
 the source domain are extended to the wild. Extensive experiments on five diffe
rent datasets validate the effectiveness of our WildNet, and we significantly ou
tperform state-of-the-art methods. The source code and model are available onlin
e: https://github.com/suhyeonlee/WildNet.
****

Auditing Privacy Defenses in Federated Learning via Generative Gradient Leakage
Zhuohang Li, Jiaxin Zhang, Luyang Liu, Jian Liu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10132-10142
Federated Learning (FL) framework brings privacy benefits to distributed learnin
g systems by allowing multiple clients to participate in a learning task under t
he coordination of a central server without exchanging their private data. Howev
er, recent studies have revealed that private information can still be leaked th
rough shared gradient information. To further protect user's privacy, several de
fense mechanisms have been proposed to prevent privacy leakage via gradient info
rmation degradation methods, such as using additive noise or gradient compressio
n before sharing it with the server. In this work, we validate that the private
training data can still be leaked under certain defense settings with a new type
 of leakage, i.e., Generative Gradient Leakage (GGL). Unlike existing methods th
at only rely on gradient information to reconstruct data, our method leverages t

he latent space of generative adversarial networks (GAN) learned from public ima ge datasets as a prior to compensate for the informational loss during gradient degradation. To address the nonlinearity caused by the gradient operator and the GAN model, we explore various gradient-free optimization methods (e.g., evoluti on strategies and Bayesian optimization) and empirically show their superiority in reconstructing high-quality images from gradients compared to gradient-based optimizers. We hope the proposed method can serve as a tool for empirically meas uring the amount of privacy leakage to facilitate the design of more robust defe nse mechanisms.

****

DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo , Hanyu Li, Xing Hu, Jirui Yuan, Zaiqing Nie; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21361-21370

Autonomous driving faces great safety challenges for a lack of global perspectiv e and the limitation of long-range perception capabilities. It has been widely a greed that vehicle-infrastructure cooperation is required to achieve Level 5 aut onomy. However, there is still NO dataset from real scenarios available for comp uter vision researchers to work on vehicle-infrastructure cooperation-related pr oblems. To accelerate computer vision research and innovation for Vehicle-Infras tructure Cooperative Autonomous Driving (VICAD), we release DAIR-V2X Dataset, wh ich is the first large-scale, multi-modal, multi-view dataset from real scenario s for VICAD. DAIR-V2X comprises 71254 LiDAR frames and 71254 Camera frames, and all frames are captured from real scenes with 3D annotations. The Vehicle-Infras tructure Cooperative 3D Object Detection problem (VIC3D) is introduced, formulat ing the problem of collaboratively locating and identifying 3D objects using sen sory input from both vehicles and infrastructure. In addition to solving traditi onal 3D object detection problems, the solution of VIC3D needs to consider the t ime asynchrony problem between vehicle and infrastructure sensors and the data t ransmission cost between them. Furthermore, we propose Time Compensation Late Fu sion (TCLF), a late fusion framework for the VIC3D task as a benchmark based on DAIR-V2X. Find data, code, and more up-to-date information at \href https://thud air.baai.ac.cn/index https://thudair.baai.ac.cn/index and \href https://github .com/AIR-THU/DAIR-V2X https://github.com/AIR-THU/DAIR-V2X .

****

DECORE: Deep Compression With Reinforcement Learning

Manoj Alwani, Yang Wang, Vashisht Madhavan; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12349-12359

Deep learning has become an increasingly popular and powerful methodology for mo dern pattern recognition systems. However, many deep neural networks have millio ns or billions of parameters, making them untenable for real-world applications due to constraints on memory size or latency requirements. As a result, efficien t network compression techniques are often required for a widespread adoption of deep learning methods. We present DECORE, a reinforcement learning based approa ch to automate the network compression process. DECORE assigns an agent to each channel in the network along with a light policy gradient method to learn which neurons or channels to be kept or removed. Each agent in the network has just on e parameter (keep or drop) to learn, which leads to a much faster training proce ss compared to existing approaches. DECORE also gives state-of-the-art compressi on results on various network architectures and various datasets. For example, o n the ResNet-110 architecture, DECORE achieves a 64.8% compression rate and 61.8 % FLOPs reduction as compared to the baseline model without any major accuracy l oss on the CIFAR-10 dataset. It can reduce the size of regular architectures lik e the VGG network by up to 99% with just a small accuracy drop of 2.28%. For a l arger dataset like ImageNet it can compress the ResNet-50 architecture by 44.7% and reduces FLOPs by 42.3%, with just a 0.69% drop on Top-5 accuracy of the unco mpressed model. We also demonstrate that DECORE can be used to search for compre ssed network architectures based on various constraints, such as memory and FLOP s.

****
Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving
Peixuan Li, Jieyu Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3885-3894
While separately leveraging monocular 3D object detection and 2D multi-object tracking can be straightforwardly applied to sequence images in a frame-by-frame fashion, stand-alone tracker cuts off the transmission of the uncertainty from the 3D detector to tracking while cannot pass tracking error differentials back to the 3D detector. In this work, we propose jointly training 3D detection and 3D tracking from only monocular videos in an end-to-end manner. The key component is a novel spatial-temporal information flow module that aggregates geometric and appearance features to predict robust similarity scores across all objects in current and past frames. Specifically, we leverage the attention mechanism of the transformer, in which self-attention aggregates the spatial information in a specific frame, and cross-attention exploits relation and affinities of all objects in the temporal domain of sequence frames. The affinities are then supervised to estimate the trajectory and guide the flow of information between corresponding 3D objects. In addition, we propose a temporal-consistency loss that explicitly involves 3D target motion modeling into the learning, making the 3D trajectory smooth in the world coordinate system. Time3D achieves 21.4% AMOTA, 13.6% AMOTP on the nuScenes 3D tracking benchmark, surpassing all published competitors, and running at 38 FPS, while Time3D achieves 31.2% mAP, 39.4% NDS on the nuScenes 3D detection benchmark.
****
MonoJSG: Joint Semantic and Geometric Cost Volume for Monocular 3D Object Detection
Qing Lian, Peiliang Li, Xiaozhi Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1070-1079
Due to the inherent ill-posed nature of 2D-3D projection, monocular 3D object detection lacks accurate depth recovery ability. Although the deep neural network (DNN) enables monocular depth-sensing from high-level learned features, the pixel-level cues are usually omitted due to the deep convolution mechanism. To benefit from both the powerful feature representation in DNN and pixel-level geometric constraints, we reformulate the monocular object depth estimation as a progressive refinement problem and propose a joint semantic and geometric cost volume to model the depth error. Specifically, we first leverage neural networks to learn the object position, dimension, and dense normalized 3D object coordinates. Based on the object depth, the dense coordinates patch together with the corresponding object features is reprojected to the image space to build a cost volume in a joint semantic and geometric error manner. The final depth is obtained by feeding the cost volume to a refinement network, where the distribution of semantic and geometric error is regularized by direct depth supervision. Through effectively mitigating depth error by the refinement framework, we achieve state-of-the-art results on both the KITTI and Waymo datasets.
****
Task Discrepancy Maximization for Fine-Grained Few-Shot Classification
SuBeen Lee, WonJun Moon, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5331-5340
Recognizing discriminative details such as eyes and beaks is important for distinguishing fine-grained classes since they have similar overall appearances. In this regard, we introduce Task Discrepancy Maximization (TDM), a simple module for fine-grained few-shot classification. Our objective is to localize the class-wise discriminative regions by highlighting channels encoding distinct information of the class. Specifically, TDM learns task-specific channel weights based on two novel components: Support Attention Module (SAM) and Query Attention Module (QAM). SAM produces a support weight to represent channel-wise discriminative power for each class. Still, since the SAM is basically only based on the labeled support sets, it can be vulnerable to bias toward such support set. Therefore, we propose QAM which complements SAM by yielding a query weight that grants more

weight to object-relevant channels for a given query image. By combining these t
wo weights, a class-wise task-specific channel weight is defined. The weights ar
e then applied to produce task-adaptive feature maps more focusing on the discri
minative details. Our experiments validate the effectiveness of TDM and its comp
lementary benefits with prior methods in fine-grained few-shot classification.
****

FedDC: Federated Learning With Non-IID Data via Local Drift Decoupling and Corre
ction

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, Cheng-Zhong Xu; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 10112-10121

Federated learning (FL) allows multiple clients to collectively train a high-per
formance global model without sharing their private data. However, the key chall
enge in federated learning is that the clients have significant statistical hete
rogeneity among their local data distributions, which would cause inconsistent o
ptimized local models on the client-side. To address this fundamental dilemma, w
e propose a novel federated learning algorithm with local drift decoupling and c
orrection (FedDC). Our FedDC only introduces lightweight modifications in the lo
cal training phase, in which each client utilizes an auxiliary local drift varia
ble to track the gap between the local model parameter and the global model para
meters. The key idea of FedDC is to utilize this learned local drift variable to
 bridge the gap, i.e., conducting consistency in parameter-level. The experiment
 results and analysis demonstrate that FedDC yields expediting convergence and b
etter performance on various image classification tasks, robust in partial parti
cipation settings, non-iid data, and heterogeneous clients.
****

Efficient Classification of Very Large Images With Tiny Objects

Fanjie Kong, Ricardo Henao; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2022, pp. 2384-2394

An increasing number of applications in computer vision, specially, in medical i
maging and remote sensing, become challenging when the goal is to classify very
large images with tiny informative objects. Specifically, these classification t
asks face two key challenges: i) the size of the input image is usually in the o
rder of mega- or giga-pixels, however, existing deep architectures do not easily
 operate on such big images due to memory constraints, consequently, we seek a m
emory-efficient method to process these images; and ii) only a very small fracti
on of the input images are informative of the label of interest, resulting in lo
w region of interest (ROI) to image ratio. However, most of the current convolut
ional neural networks (CNNs) are designed for image classification datasets that
 have relatively large ROIs and small image sizes (sub-megapixel). Existing appr
oaches have addressed these two challenges in isolation. We present an end-to-en
d CNN model termed Zoom-In network that leverages hierarchical attention samplin
g for classification of large images with tiny objects using a single GPU. We ev
aluate our method on four large-image histopathology, road-scene and satellite i
maging datasets, and one gigapixel pathology dataset. Experimental results show
that our model achieves higher accuracy than existing methods while requiring le
ss memory resources.
****

SWEM: Towards Real-Time Video Object Segmentation With Sequential Weighted Expec
tation-Maximization

Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, Wei Liu;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2022, pp. 1362-1372

Matching-based methods, especially those based on space-time memory, are signifi
cantly ahead of other solutions in semi-supervised video object segmentation (VO
S). However, continuously growing and redundant template features lead to an ine
fficient inference. To alleviate this, we propose a novel Sequential Weighted Ex
pectation-Maximization (SWEM) network to greatly reduce the redundancy of memory
 features. Different from the previous methods which only detect feature redunda
ncy between frames, SWEM merges both intra-frame and inter-frame similar feature

s by leveraging the sequential weighted EM algorithm. Further, adaptive weights for frame features endow SWEM with the flexibility to represent hard samples, improving the discrimination of templates. Besides, the proposed method maintains a fixed number of template features in memory, which ensures the stable inference complexity of the VOS system. Extensive experiments on commonly used DAVIS and YouTube-VOS datasets verify the high efficiency (36 FPS) and high performance (84.3% J&F on DAVIS 2017 validation dataset) of SWEM.
****

Point-to-Voxel Knowledge Distillation for LiDAR Semantic Segmentation
Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, Yikang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8479-8488
This article addresses the problem of distilling knowledge from a large teacher model to a slim student network for LiDAR semantic segmentation. Directly employing previous distillation approaches yields inferior results due to the intrinsic challenges of point cloud, i.e., sparsity, randomness and varying density. To tackle the aforementioned problems, we propose the Point-to-Voxel Knowledge Distillation (PVD), which transfers the hidden knowledge from both point level and voxel level. Specifically, we first leverage both the pointwise and voxelwise output distillation to complement the sparse supervision signals. Then, to better exploit the structural information, we divide the whole point cloud into several supervoxels and design a difficultyaware sampling strategy to more frequently sample supervoxels containing less frequent classes and faraway objects. On these supervoxels, we propose inter-point and intervoxel affinity distillation, where the similarity information between points and voxels can help the student model better capture the structural information of the surrounding environment. We conduct extensive experiments on two popular LiDAR segmentation benchmarks, i.e., nuScenes [3] and SemanticKITTI [1]. On both benchmarks, our PVD consistently outperforms previous distillation approaches by a large margin on three representative backbones, i.e., Cylinder3D [27, 28], SPVNAS [20] and MinkowskiNet [5]. Notably, on the challenging nuScenes and SemanticKITTI datasets, our method can achieve roughly 75% MACs reduction and 2x speedup on the competitive Cylinder3D model and rank 1st on the SemanticKITTI leaderboard among all published algorithms. Our code is available at https://github.com/cardwing/Codes-for-PVKD.
****

Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers
Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, Chris Russell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10410-10421
Algorithmic fairness is frequently motivated in terms of a trade-off in which overall performance is decreased so as to improve performance on disadvantaged groups where the algorithm would otherwise be less accurate. Contrary to this, we find that applying existing fairness approaches to computer vision improve fairness by degrading the performance of classifiers across all groups (with increased degradation on the best performing groups). Extending the bias-variance decomposition for classification to fairness, we theoretically explain why the majority of fairness methods designed for low capacity models should not be used in settings involving high-capacity models, a scenario common to computer vision. We corroborate this analysis with extensive experimental support that shows that many of the fairness heuristics used in computer vision also degrade performance on the most disadvantaged groups. Building on these insights, we propose an adaptive augmentation strategy that, uniquely, of all methods tested, improves performance for the disadvantaged groups.
****

Generating Diverse 3D Reconstructions From a Single Occluded Face Image
Rahul Dey, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1547-1557
Occlusions are a common occurrence in unconstrained face images. Single image 3D reconstruction from such face images often suffers from corruption due to the p

resence of occlusions. Furthermore, while a plurality of 3D reconstructions is p
lausible in the occluded regions, existing approaches are limited to generating
only a single solution. To address both of these challenges, we present Diverse3
DFace, which is specifically designed to simultaneously generate a diverse and r
ealistic set of 3D reconstructions from a single occluded face image. It consist
s of three components: a global+local shape fitting process, a graph neural netw
ork-based mesh VAE, and a Determinantal Point Process based diversity promoting
iterative optimization procedure. Quantitative and qualitative comparisons of 3D
 reconstruction on occluded faces show that Diverse3DFace can estimate 3D shapes
 that are consistent with the visible regions in the target image while exhibiti
ng high, yet realistic, levels of diversity on the occluded regions. On face ima
ges occluded by masks, glasses, and other random objects, Diverse3DFace generate
s a distribution of 3D shapes having 50% higher diversity on the occluded regio
ns compared to the baselines. Moreover, our closest sample to the ground truth h
as 40% lower MSE than the singular reconstructions by existing approaches. Code
 and data available at: https://github.com/human-analysis/diverse3dface
****

RBGNet: Ray-Based Grouping for 3D Object Detection
Haiyang Wang, Shaoshuai Shi, Ze Yang, Rongyao Fang, Qi Qian, Hongsheng Li, Bernt
 Schiele, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2022, pp. 1110-1119
As a fundamental problem in computer vision, 3D object detection is experiencing
 rapid growth. To extract the point-wise features from the irregularly and spars
ely distributed points, previous methods usually take a feature grouping module
to aggregate the point features to an object candidate. However, these methods h
ave not yet leveraged the surface geometry of foreground objects to enhance grou
ping and 3D box generation. In this paper, we propose the RBGNet framework, a vo
ting-based 3D detector for accurate 3D object detection from point clouds. In or
der to learn better representations of object shape to enhance cluster features
for predicting 3D boxes, we propose a ray-based feature grouping module, which a
ggregates the point-wise features on object surfaces using a group of determined
 rays uniformly emitted from cluster centers. Considering the fact that foregrou
nd points are more meaningful for box estimation, we design a novel foreground b
iased sampling strategy in downsample process to sample more points on object su
rfaces and further boost the detection performance. Our model achieves state-of-
the-art 3D detection performance on ScanNet V2 and SUN RGB-D with remarkable per
formance gains.
****

Stand-Alone Inter-Frame Attention in Video Models
Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Jiebo Luo, Tao Mei; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
022, pp. 3192-3201
Motion, as the uniqueness of a video, has been critical to the development of vi
deo understanding models. Modern deep learning models leverage motion by either
executing spatio-temporal 3D convolutions, factorizing 3D convolutions into spat
ial and temporal convolutions separately, or computing self-attention along temp
oral dimension. The implicit assumption behind such successes is that the featur
e maps across consecutive frames can be nicely aggregated. Nevertheless, the ass
umption may not always hold especially for the regions with large deformation. I
n this paper, we present a new recipe of inter-frame attention block, namely Sta
nd-alone Inter-Frame Attention (SIFA), that novelly delves into the deformation
across frames to estimate local self-attention on each spatial location. Technic
ally, SIFA remoulds the deformable design via re-scaling the offset predictions
by the difference between two frames. Taking each spatial location in the curren
t frame as the query, the locally deformable neighbors in the next frame are reg
arded as the keys/values. Then, SIFA measures the similarity between query and k
eys as stand-alone attention to weighted average the values for temporal aggrega
tion. We further plug SIFA block into ConvNets and Vision Transformer, respectiv
ely, to devise SIFA-Net and SIFA-Transformer. Extensive experiments conducted on
 four video datasets demonstrate the superiority of SIFA-Net and SIFA-Transforme

r as stronger backbones. More remarkably, SIFA-Transformer achieves an accuracy of 83.1% on Kinetics-400 dataset. Source code is available at https://github.com/FuchenUSTC/SIFA.

****

Uncertainty-Aware Adaptation for Self-Supervised 3D Human Pose Estimation

Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20448-20459

The advances in monocular 3D human pose estimation are dominated by supervised techniques that require large-scale 2D/3D pose annotations. Such methods often behave erratically in the absence of any provision to discard unfamiliar out-of-distribution data. To this end, we cast the 3D human pose learning as an unsupervised domain adaptation problem. We introduce MRP-Net that constitutes a common deep network backbone with two output heads subscribing to two diverse configurations; a) model-free joint localization and b) model-based parametric regression. Such a design allows us to derive suitable measures to quantify prediction uncertainty at both pose and joint level granularity. While supervising only on labeled synthetic samples, the adaptation process aims to minimize the uncertainty for the unlabeled target images while maximizing the same for an extreme out-of-distribution dataset (backgrounds). Alongside synthetic-to-real 3D pose adaptation, the joint-uncertainties allow expanding the adaptation to work on in-the-wild images even in the presence of occlusion and truncation scenarios. We present a comprehensive evaluation of the proposed approach and demonstrate state-of-the-art performance on benchmark datasets.

****

Open-Domain, Content-Based, Multi-Modal Fact-Checking of Out-of-Context Images via Online Resources

Sahar Abdelnabi, Rakibul Hasan, Mario Fritz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14940-14949

Misinformation is now a major problem due to its potential high risks to our core democratic and societal values and orders. Out-of-context misinformation is one of the easiest and effective ways used by adversaries to spread viral false stories. In this threat, a real image is re-purposed to support other narratives by misrepresenting its context and/or elements. The internet is being used as the go-to way to verify information using different sources and modalities. Our goal is an inspectable method that automates this time-consuming and reasoning-intensive process by fact-checking the image-caption pairing using Web evidence. To integrate evidence and cues from both modalities, we introduce the concept of 'multi-modal cycle-consistency check'; starting from the image/caption, we gather textual/visual evidence, which will be compared against the other paired caption/image, respectively. Moreover, we propose a novel architecture, Consistency-Checking Network (CCN), that mimics the layered human reasoning across the same and different modalities: the caption vs. textual evidence, the image vs. visual evidence, and the image vs. caption. Our work offers the first step and benchmark for open-domain, content-based, multi-modal fact-checking, and significantly outperforms previous baselines that did not leverage external evidence.

****

Memory-Augmented Deep Conditional Unfolding Network for Pan-Sharpening

Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, Fan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1788-1797

Pan-sharpening aims to obtain high-resolution multispectral (MS) images for remote sensing systems and deep learning-based methods have achieved remarkable success. However, most existing methods are designed in a black-box principle, lacking sufficient interpretability. Additionally, they ignore the different characteristics of each band of MS images and directly concatenate them with panchromatic (PAN) images, leading to severe copy artifacts. To address the above issues, we propose an interpretable deep neural network, namely Memory-augmented Deep Conditional Unfolding Network with two specified core designs. Firstly, considering the degradation process, it formulates the Pan-sharpening problem as the minimi

zation of a variational model with denoising-based prior and non-local auto-regr
ession prior which is capable of searching the similarities between long-range p
atches, benefiting the texture enhancement. A novel iteration algorithm with bui
lt-in CNNs is exploited for transparent model design. Secondly, to fully explore
 the potentials of different bands of MS images, the PAN image is combined with
each band of MS images, selectively providing the high-frequency details and all
eviating the copy artifacts. Extensive experimental results validate the superio
rity of the proposed algorithm against other state-of-the-art methods.
****

Semi-Supervised Wide-Angle Portraits Correction by Multi-Scale Transformer
Fushun Zhu, Shan Zhao, Peng Wang, Hao Wang, Hua Yan, Shuaicheng Liu; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
022, pp. 19689-19698
We propose a semi-supervised network for wide-angle portraits correction. Wide-a
ngle images often suffer from skew and distortion affected by perspective distor
tion, especially noticeable at the face regions. Previous deep learning based ap
proaches need the ground-truth correction flow maps for training guidance. Howev
er, such labels are expensive, which can only be obtained manually. In this work
, we design a semi-supervised scheme and build a high-quality unlabeled dataset
with rich scenarios, allowing us to simultaneously use labeled and unlabeled dat
a to improve performance. Specifically, our semi-supervised scheme takes advanta
ge of the consistency mechanism, with several novel components such as direction
 and range consistency (DRC) and regression consistency (RC). Furthermore, diffe
rent from the existing methods, we propose the Multi-Scale Swin-Unet (MS-Unet) b
ased on the multi-scale swin transformer block (MSTB), which can simultaneously
learn short-distance and long-distance information to avoid artifacts. Extensive
 experiments demonstrate that the proposed method is superior to the state-of-th
e-art methods and other representative baselines.
****

Large-Scale Pre-Training for Person Re-Identification With Noisy Labels
Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang L
i, Fang Wen, Dong Chen; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2022, pp. 2476-2486
This paper aims to address the problem of pre-training for person re-identificat
ion (Re-ID) with noisy labels. To setup the pre-training task, we apply a simple
 online multi-object tracking system on raw videos of an existing unlabeled Re-I
D dataset "LUPerson" and build the Noisy Labeled variant called "LUPerson-NL". S
ince theses ID labels automatically derived from tracklets inevitably contain no
ises, we develop a large-scale Pre-training framework utilizing Noisy Labels (PN
L), which consists of three learning modules: supervised Re-ID learning, prototy
pe-based contrastive learning, and label-guided contrastive learning. In princip
le, joint learning of these three modules not only clusters similar examples to
one prototype, but also rectifies noisy labels based on the prototype assignment
. We demonstrate that learning directly from raw videos is a promising alternati
ve for pre-training, which utilizes spatial and temporal correlations as weak su
pervision. This simple pre-training task provides a scalable way to learn SOTA R
e-ID representations from scratch on "LUPerson-NL" without bells and whistles. F
or example, by applying on the same supervised Re-ID method MGN, our pre-trained
 model improves the mAP over the unsupervised pre-training counterpart by 5.7%,
2.2%, 2.3% on CUHK03, DukeMTMC, and MSMT17 respectively. Under the small-scale
or few-shot setting, the performance gain is even more significant, suggesting a
 better transferability of the learned representation. Code is available at htt
ps://github.com/DengpanFu/LUPerson-NL
****

Adiabatic Quantum Computing for Multi Object Tracking
Jan-Nico Zaech, Alexander Liniger, Martin Danelljan, Dengxin Dai, Luc Van Gool;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2022, pp. 8811-8822
Multi-Object Tracking (MOT) is most often approached in the tracking-by-detectio
n paradigm, where object detections are associated through time. The association

step naturally leads to discrete optimization problems. As these optimization problems are often NP-hard, they can only be solved exactly for small instances on current hardware. Adiabatic quantum computing (AQC) offers a solution for this, as it has the potential to provide a considerable speedup on a range of NP-hard optimization problems in the near future. However, current MOT formulations are unsuitable for quantum computing due to their scaling properties. In this work, we therefore propose the first MOT formulation designed to be solved with AQC. We employ an Ising model that represents the quantum mechanical system implemented on the AQC. We show that our approach is competitive compared with state-of-the-art optimization-based approaches, even when using of-the-shelf integer programming solvers. Finally, we demonstrate that our MOT problem is already solvable on the current generation of real quantum computers for small examples, and analyze the properties of the measured solutions.

****

Feature Erasing and Diffusion Network for Occluded Person Re-Identification
Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, Jiangning Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4754-4763

Occluded person re-identification (ReID) aims at matching occluded person images to holistic ones across different camera views. Target Pedestrians (TP) are often disturbed by Non-Pedestrian Occlusions (NPO) and Non-Target Pedestrians (NTP). Previous methods mainly focus on increasing the model's robustness against NPO while ignoring feature contamination from NTP. In this paper, we propose a novel Feature Erasing and Diffusion Network (FED) to simultaneously handle challenges from NPO and NTP. Specifically, aided by the NPO augmentation strategy that simulates NPO on holistic pedestrian images and generates precise occlusion masks, NPO features are explicitly eliminated by our proposed Occlusion Erasing Module (OEM). Subsequently, we diffuse the pedestrian representations with other memorized features to synthesize the NTP characteristics in the feature space through the novel Feature Diffusion Module (FDM). With the guidance of the occlusion scores from OEM, the feature diffusion process is conducted on visible body parts, thereby improving the quality of the synthesized NTP characteristics. We can greatly improve the model's perception ability towards TP and alleviate the influence of NPO and NTP by jointly optimizing OEM and FDM. Furthermore, the proposed FDM works as an auxiliary module for training and will not be engaged in the inference phase, thus with high flexibility. Experiments on occluded and holistic person ReID benchmarks demonstrate the superiority of FED over state-of-the-art methods.

****

Is Mapping Necessary for Realistic PointGoal Navigation?
Ruslan Partsey, Erik Wijmans, Naoki Yokoyama, Oles Dobosevych, Dhruv Batra, Oleksandr Maksymets; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17232-17241

Can an autonomous agent navigate in a new environment without building an explicit map? For the task of PointGoal navigation ('Go to (x, y)') under idealized settings (no RGB-D and actuation noise, perfect GPS+Compass), the answer is a clear 'yes' - map-less neural models composed of task-agnostic components (CNNs and RNNs) trained with large-scale reinforcement learning achieve 100% Success on a standard dataset (Gibson). However, for PointNav in a realistic setting (RGB-D and actuation noise, no GPS+Compass), this is an open question; one we tackle in this paper. The strongest published result for this task is 71.7% Success. First, we identify the main (perhaps, only) cause of the drop in performance: absence of GPS+Compass. An agent with perfect GPS+Compass faced with RGB-D sensing and actuation noise achieves 99.8% Success (Gibson-v2 val). This suggests that (to paraphrase a meme) robust visual odometry is all we need for realistic PointNav; if we can achieve that, we can ignore the sensing and actuation noise. With that as our operating hypothesis, we scale dataset size, model size, and develop human-annotation-free data-augmentation techniques to train neural models for visual odometry. We advance state of the art on the Habitat Realistic PointNav Challenge - SPL by 40% (relative), 53 to 74, and Success by 31% (relative), 71 to 94.

While our approach does not saturate or 'solve' this dataset, this strong improvement combined with promising zero-shot sim2real transfer (to a LoCoBot robot) provides evidence consistent with the hypothesis that explicit mapping may not be necessary for navigation, even in realistic setting.
****

Node-Aligned Graph Convolutional Network for Whole-Slide Image Representation and Classification
Yonghang Guan, Jun Zhang, Kuan Tian, Sen Yang, Pei Dong, Jinxi Xiang, Wei Yang, Junzhou Huang, Yuyao Zhang, Xiao Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18813-18823
The large-scale whole-slide images (WSIs) facilitate the learning-based computational pathology methods. However, the gigapixel size of WSIs makes it hard to train a conventional model directly. Current approaches typically adopt multiple-instance learning (MIL) to tackle this problem. Among them, MIL combined with graph convolutional network (GCN) is a significant branch, where the sampled patches are regarded as the graph nodes to further discover their correlations. However, it is difficult to build correspondence across patches from different WSIs. Therefore, most methods have to perform non-ordered node pooling to generate the bag-level representation. Direct non-ordered pooling will lose much structural and contextual information, such as patch distribution and heterogeneous patterns, which is critical for WSI representation. In this paper, we propose a hierarchical global-to-local clustering strategy to build a Node-Aligned GCN (NAGCN) to represent WSI with rich local structural information as well as global distribution. We first deploy a global clustering operation based on the instance features in the dataset to build the correspondence across different WSIs. Then, we perform a local clustering-based sampling strategy to select typical instances belonging to each cluster within the WSI. Finally, we employ the graph convolution to obtain the representation. Since our graph construction strategy ensures the alignment among different WSIs, WSI-level representation can be easily generated and used for the subsequent classification. The experiment results on two cancer subtype classification datasets demonstrate our method achieves better performance compared with the state-of-the-art methods.
****

Represent, Compare, and Learn: A Similarity-Aware Framework for Class-Agnostic Counting
Min Shi, Hao Lu, Chen Feng, Chengxin Liu, Zhiguo Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9529-9538
Class-agnostic counting (CAC) aims to count all instances in a query image given few exemplars. A standard pipeline is to extract visual features from exemplars and match them with query images to infer object counts. Two essential components in this pipeline are feature representation and similarity metric. Existing methods either adopt a pretrained network to represent features or learn a new one, while applying a naive similarity metric with fixed inner product. We find this paradigm leads to noisy similarity matching and hence harms counting performance. In this work, we propose a similarity-aware CAC framework that jointly learns representation and similarity metric. We first instantiate our framework with a naive baseline called Bilinear Matching Network (BMNet), whose key component is a learnable bilinear similarity metric. To further embody the core of our framework, we extend BMNet to BMNet+ that models similarity from three aspects: 1) representing the instances via their self-similarity to enhance feature robustness against intra-class variations; 2) comparing the similarity dynamically to focus on the key patterns of each exemplar; 3) learning from a supervision signal to impose explicit constraints on matching results. Extensive experiments on a recent CAC dataset FSC147 show that our models significantly outperform state-of-the-art CAC approaches. In addition, we also validate the cross-dataset generality of BMNet and BMNet+ on a car counting dataset CARPK. Code is at tiny.one/BMNet
****

Masked Feature Prediction for Self-Supervised Visual Pre-Training

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, Christoph Feichtenhofer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14668-14678

We present Masked Feature Prediction (MaskFeat) for self-supervised pre-training of video models. Our approach first randomly masks out a portion of the input sequence and then predicts the feature of the masked regions. We study five different types of features and find Histograms of Oriented Gradients (HOG), a hand-crafted feature descriptor, works particularly well in terms of both performance and efficiency. We observe that the local contrast normalization in HOG is essential for good results, which is in line with earlier work using HOG for visual recognition. Our approach can learn abundant visual knowledge and drive large-scale Transformer-based models. Without using extra model weights or supervision, MaskFeat pre-trained on unlabeled videos achieves unprecedented results of 86.7% with MViTv2-L on Kinetics-400, 88.3% on Kinetics-600, 80.4% on Kinetics-700, 38.8 mAP on AVA, and 75.0% on SSv2. MaskFeat further generalizes to image input, which can be interpreted as a video with a single frame and obtains competitive results on ImageNet.

****

Critical Regularizations for Neural Surface Reconstruction in the Wild

Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6270-6279

Neural implicit functions have recently shown promising results on surface reconstructions from multiple views. However, current methods still suffer from excessive time complexity and poor robustness when reconstructing unbounded or complex scenes. In this paper, we present RegSDF, which shows that proper point cloud supervisions and geometry regularizations are sufficient to produce high-quality and robust reconstruction results. Specifically, RegSDF takes an additional oriented point cloud as input, and optimizes a signed distance field and a surface light field within a differentiable rendering framework. We also introduce the two critical regularizations for this optimization. The first one is the Hessian regularization that smoothly diffuses the signed distance values to the entire distance field given noisy and incomplete input. And the second one is the minimal surface regularization that compactly interpolates and extrapolates the missing geometry. Extensive experiments are conducted on DTU, BlendedMVS, and Tanks and Temples datasets. Compared with recent neural surface reconstruction approaches, RegSDF is able to reconstruct surfaces with fine details even for open scenes with complex topologies and unstructured camera trajectories.

****

EASE: Unsupervised Discriminant Subspace Learning for Transductive Few-Shot Learning

Hao Zhu, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9078-9088

Few-shot learning (FSL) has received a lot of attention due to its remarkable ability to adapt to novel classes. Although many techniques have been proposed for FSL, they mostly focus on improving FSL backbones. Some works also focus on learning on top of the features generated by these backbones to adapt them to novel classes. We present an unsupErvised discriminAnt Subspace lEarning (EASE) that improves transductive few-shot learning performance by learning a linear projection onto a subspace built from features of the support set and the unlabeled query set in the test time. Specifically, based on the support set and the unlabeled query set, we generate the similarity matrix and the dissimilarity matrix based on the structure prior for the proposed EASE method, which is efficiently solved with SVD. We also introduce conStraIned wAsserstein MEan Shift clustEring (SIAMESE) which extends Sinkhorn K-means by incorporating labeled support samples. SIAMESE works on the features obtained from EASE to estimate class centers and query predictions. On the mini-ImageNet, tiered-ImageNet, CIFAR-FS, CUB and OpenMIC benchmarks, both steps significantly boost the performance in transductive FSL and semi-supervised FSL.

****

Object-Relation Reasoning Graph for Action Recognition
Yangjun Ou, Li Mi, Zhenzhong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20133-20142
Action recognition is a challenging task since the attributes of objects as well as their relationships change constantly in the video. Existing methods mainly use object-level graphs or scene graphs to represent the dynamics of objects and relationships, but ignore modeling the fine-grained relationship transitions directly. In this paper, we propose an Object-Relation Reasoning Graph (OR2G) for reasoning about action in videos. By combining an object-level graph (OG) and a relation-level graph (RG), the proposed OR2G catches the attribute transitions of objects and reasons about the relationship transitions between objects simultaneously. In addition, a graph aggregating module (GAM) is investigated by applying the multi-head edge-to-node message passing operation. GAM feeds back the information from the relation node to the object node and enhances the coupling between the object-level graph and the relation-level graph. Experiments in video action recognition demonstrate the effectiveness of our approach when compared with the state-of-the-art methods.
****

Semantic Segmentation by Early Region Proxy
Yifan Zhang, Bo Pang, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1258-1268
Typical vision backbones manipulate structured features. As a compromise, semantic segmentation has long been modeled as per-point prediction on dense regular grids. In this work, we present a novel and efficient modeling that starts from interpreting the image as a tessellation of learnable regions, each of which has flexible geometrics and carries homogeneous semantics. To model region-wise context, we exploit Transformer to encode regions in a sequence-to-sequence manner by applying multi-layer self-attention on the region embeddings, which serve as proxies of specific regions. Semantic segmentation is now carried out as per-region prediction on top of the encoded region embeddings using a single linear classifier, where a decoder is no longer needed. The proposed RegProxy model discards the common Cartesian feature layout and operates purely at region level. Hence, it exhibits the most competitive performance-efficiency trade-off compared with the conventional dense prediction methods. For example, on ADE20K, the small-sized RegProxy-S/16 outperforms the best CNN model using 25% parameters and 4% computation, while the largest RegProxy-L/16 achieves 52.9mIoU which outperforms the state-of-the-art by 2.1% with fewer resources. Codes and models are available at https://github.com/YiF-Zhang/RegionProxy.
****

GIQE: Generic Image Quality Enhancement via Nth Order Iterative Degradation
Pranjay Shyam, Kyung-Soo Kim, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2077-2087
Visual degradations caused by motion blur, raindrop, rain, snow, illumination, and fog deteriorate image quality and, subsequently, the performance of perception algorithms deployed in outdoor conditions. While degradation-specific image restoration techniques have been extensively studied, such algorithms are domain sensitive and fail in real scenarios where multiple degradations exist simultaneously. This makes a case for blind image restoration and reconstruction algorithms as practically relevant. However, the absence of a dataset diverse enough to encapsulate all variations hinders development for such an algorithm. In this paper, we utilize a synthetic degradation model that recursively applies sets of random degradations to generate naturalistic degradation images of varying complexity, which are used as input. Furthermore, as the degradation intensity can vary across an image, the spatially invariant convolutional filter cannot be applied for all degradations. Hence to enable spatial variance during image restoration and reconstruction, we design a transformer-based architecture to benefit from the long-range dependencies. In addition, to reduce the computational cost of transformers, we propose a multi-branch structure coupled with modifications such as a complimentary feature selection mechanism and the replacement of a feed-forward network with lightweight multiscale convolutions. Finally, to improve resto

ration and reconstruction, we integrate an auxiliary decoder branch to predict t
he degradation mask to ensure the underlying network can localize the degradatio
n information. From empirical analysis on 10 datasets covering rain drop removal
, deraining, dehazing, image enhancement, and deblurring, we demonstrate the eff
icacy of the proposed approach while obtaining SoTA performance.
****

Instance Segmentation With Mask-Supervised Polygonal Boundary Transformers
Justin Lazarow, Weijian Xu, Zhuowen Tu; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4382-4391
In this paper, we present an end-to-end instance segmentation method that regres
ses a polygonal boundary for each object instance. This sparse, vectorized bound
ary representation for objects, while attractive in many downstream computer vis
ion tasks, quickly runs into issues of parity that need to be addressed: parity
in supervision and parity in performance when compared to existing pixel-based m
ethods. This is due in part to object instances being annotated with ground-trut
h in the form of polygonal boundaries or segmentation masks, yet being evaluated
 in a convenient manner using only segmentation masks. Our method, named Boundar
yFormer, is a Transformer based architecture that directly predicts polygons yet
 uses instance mask segmentations as the ground-truth supervision for computing
the loss. We achieve this by developing an end-to-end differentiable model that
solely relies on supervision within the mask space through differentiable raster
ization. BoundaryFormer matches or surpasses the Mask R-CNN method in terms of i
nstance segmentation quality on both COCO and Cityscapes while exhibiting signif
icantly better transferability across datasets.
****

FaceVerse: A Fine-Grained and Detail-Controllable 3D Face Morphable Model From a
 Hybrid Dataset
Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, Yebin Liu; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2022, pp. 20333-20342
We present FaceVerse, a fine-grained 3D Neural Face Model, which is built from h
ybrid East Asian face datasets containing 60K fused RGB-D images and 2K high-fid
elity 3D head scan models. A novel coarse-to-fine structure is proposed to take
better advantage of our hybrid dataset. In the coarse module, we generate a base
 parametric model from large-scale RGB-D images, which is able to predict accura
te rough 3D face models in different genders, ages, etc. Then in the fine module
, a conditional StyleGAN architecture trained with high-fidelity scan models is
introduced to enrich elaborate facial geometric and texture details. Note that d
ifferent from previous methods, our base and detailed modules are both changeabl
e, which enables an innovative application of adjusting both the basic attribute
s and the facial details of 3D face models. Furthermore, we propose a single-ima
ge fitting framework based on differentiable rendering. Rich experiments show th
at our method outperforms the state-of-the-art methods.
****

Bring Evanescent Representations to Life in Lifelong Class Incremental Learning
Marco Toldo, Mete Ozay; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2022, pp. 16732-16741
In Class Incremental Learning (CIL), a classification model is progressively tra
ined at each incremental step on an evolving dataset of new classes, while at th
e same time, it is required to preserve knowledge of all the classes observed so
 far. Prototypical representations can be leveraged to model feature distributio
n for the past data and inject information of former classes in later incrementa
l steps without resorting to stored exemplars. However, if not updated, those re
presentations become increasingly outdated as the incremental learning progresse
s with new classes. To address the aforementioned problems, we propose a framewo
rk which aims to (i) model the semantic drift by learning the relationship betwe
en representations of past and novel classes among incremental steps, and (ii) e
stimate the feature drift, defined as the evolution of the representations learn
ed by models at each incremental step. Semantic and feature drifts are then join
tly exploited to infer up-to-date representations of past classes (evanescent re

presentations), and thereby infuse past knowledge into incremental training. We experimentally evaluate our framework achieving exemplar-free SotA results on multiple benchmarks. In the ablation study, we investigate nontrivial relationships between evanescent representations and models.
****

Single-Stage 3D Geometry-Preserving Depth Estimation Model Training on Dataset Mixtures With Uncalibrated Stereo Data
Nikolay Patakin, Anna Vorontsova, Mikhail Artemyev, Anton Konushin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1705-1714
Nowadays, robotics, AR, and 3D modeling applications attract considerable attention to single-view depth estimation (SVDE) as it allows estimating scene geometry from a single RGB image. Recent works have demonstrated that the accuracy of an SVDE method hugely depends on the diversity and volume of the training data. However, RGB-D datasets obtained via depth capturing or 3D reconstruction are typically small, synthetic datasets are not photorealistic enough, and all these datasets lack diversity. The large-scale and diverse data can be sourced from stereo images or stereo videos from the web. Typically being uncalibrated, stereo data provides disparities up to unknown shift (geometrically incomplete data), so stereo-trained SVDE methods cannot recover 3D geometry. It was recently shown that the distorted point clouds obtained with a stereo-trained SVDE method can be corrected with additional point cloud modules (PCM) separately trained on the geometrically complete data. On the contrary, we propose GP2, General-Purpose and Geometry-Preserving training scheme, and show that conventional SVDE models can learn correct shifts themselves without any post-processing, benefiting from using stereo data even in the geometry-preserving setting. Through experiments on different dataset mixtures, we prove that GP2-trained models outperform methods relying on PCM in both accuracy and speed, and report the state-of-the-art results in the general-purpose geometry-preserving SVDE. Moreover, we show that SVDE models can learn to predict geometrically correct depth even when geometrically complete data comprises the minor part of the training set.
****

LD-ConGR: A Large RGB-D Video Dataset for Long-Distance Continuous Gesture Recognition
Dan Liu, Libo Zhang, Yanjun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3304-3312
Gesture recognition plays an important role in natural human-computer interaction and sign language recognition. Existing research on gesture recognition is limited to close-range interaction such as vehicle gesture control and face-to-face communication. To apply gesture recognition to long-distance interactive scenes such as meetings and smart homes, a large RGB-D video dataset LD-ConGR is established in this paper. LD-ConGR is distinguished from existing gesture datasets by its long-distance gesture collection, fine-grained annotations, and high video quality. Specifically, 1) the farthest gesture provided by the LD-ConGR is captured 4m away from the camera while existing gesture datasets collect gestures within 1m from the camera; 2) besides the gesture category, the temporal segmentation of gestures and hand location are also annotated in LD-ConGR; 3) videos are captured at high resolution (1280x720 for color streams and 640x576 for depth streams) and high frame rate (30 fps). On top of the LD-ConGR, a series of experimental and studies are conducted, and the proposed gesture region estimation and key frame sampling strategies are demonstrated to be effective in dealing with long-distance gesture recognition and the uncertainty of gesture duration. The dataset and experimental results presented in this paper are expected to boost the research of long-distance gesture recognition. The dataset is available at https://github.com/Diananini/LD-ConGR-CVPR2022.
****

SimVQA: Exploring Simulated Environments for Visual Question Answering
Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio S. Feris, Vicente Ordonez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5056-5066

Existing work on VQA explores data augmentation to achieve better generalization by perturbing the images in the dataset or modifying the existing questions and answers. While these methods exhibit good performance, the diversity of the questions and answers are constrained by the available image set. In this work we explore using synthetic computer-generated data to fully control the visual and language space, allowing us to provide more diverse scenarios. We quantify the effect of synthetic data in real-world VQA benchmarks and to which extent it produces results that generalize to real data. By exploiting 3D and physics simulation platforms, we provide a pipeline to generate synthetic data to expand and replace type-specific questions and answers without risking the exposure of sensitive or personal data that might be present in real images. We offer a comprehensive analysis while expanding existing hyper-realistic datasets to be used for VQA. We also propose Feature Swapping (F-SWAP) -- where we randomly switch object-level features during training to make a VQA model more domain invariant. We show that F-SWAP is effective for enhancing a currently existing VQA dataset of real images without compromising on the accuracy to answer existing questions in the dataset.

****

Thin-Plate Spline Motion Model for Image Animation

Jian Zhao, Hui Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3657-3666

Image animation brings life to the static object in the source image according to the driving video. Recent works attempt to perform motion transfer on arbitrary objects through unsupervised methods without using a priori knowledge. However, it remains a significant challenge for current unsupervised methods when there is a large pose gap between the objects in the source and driving images. In this paper, a new end-to-end unsupervised motion transfer framework is proposed to overcome such issue. Firstly, we propose thin-plate spline motion estimation to produce a more flexible optical flow, which warps the feature maps of the source image to the feature domain of the driving image. Secondly, in order to restore the missing regions more realistically, we leverage multi-resolution occlusion masks to achieve more effective feature fusion. Finally, additional auxiliary loss functions are designed to ensure that there is a clear division of labor in the network modules, encouraging the network to generate high-quality images. Our method can animate a variety of objects, including talking faces, human bodies, and pixel animations. Experiments demonstrate that our method performs better on most benchmarks than the state of the art with visible improvements in pose-related metrics.

****

Learning Local Displacements for Point Cloud Completion

Yida Wang, David Joseph Tan, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1568-1577

We propose a novel approach aimed at object and semantic scene completion from a partial scan represented as a 3D point cloud. Our architecture relies on three novel layers that are used successively within an encoder-decoder structure and specifically developed for the task at hand. The first one carries out feature extraction by matching the point features to a set of pre-trained local descriptors. Then, to avoid losing individual descriptors as part of standard operations such as max-pooling, we propose an alternative neighbor-pooling operation that relies on adopting the feature vectors with the highest activations. Finally, up-sampling in the decoder modifies our feature extraction in order to increase the output dimension. While this model is already able to achieve competitive results with the state of the art, we further propose a way to increase the versatility of our approach to process point clouds. To this aim, we introduce a second model that assembles our layers within a transformer architecture. We evaluate both architectures on object and indoor scene completion tasks, achieving state-of-the-art performance.

****

Human Hands As Probes for Interactive Object Understanding

Mohit Goyal, Sahil Modi, Rishabh Goyal, Saurabh Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3293-3303
Interactive object understanding, or what we can do to objects and how is a long-standing goal of computer vision. In this paper, we tackle this problem through observation of human hands in in-the-wild egocentric videos. We demonstrate that observation of what human hands interact with and how can provide both the relevant data and the necessary supervision. Attending to hands, readily localizes and stabilizes active objects for learning and reveals places where interactions with objects occur. Analyzing the hands shows what we can do to objects and how. We apply these basic principles on the EPIC-KITCHENS dataset, and successfully learn state-sensitive features, and object affordances (regions of interaction and afforded grasps), purely by observing hands in egocentric videos.
****

Understanding and Increasing Efficiency of Frank-Wolfe Adversarial Training
Theodoros Tsiligkaridis, Jay Roberts; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 50-59
Deep neural networks are easily fooled by small perturbations known as adversarial attacks. Adversarial Training (AT) is a technique that approximately solves a robust optimization problem to minimize the worst-case loss and is widely regarded as the most effective defense against such attacks. Due to the high computation time for generating strong adversarial examples in the AT process, single-step approaches have been proposed to reduce training time. However, these methods suffer from catastrophic overfitting where adversarial accuracy drops during training, and although improvements have been proposed, they increase training time and robustness is far from that of multi-step AT. We develop a theoretical framework for adversarial training with FW optimization (FW-AT) that reveals a geometric connection between the loss landscape and the distortion of l-inf FW attacks (the attack's l-2 norm). Specifically, we analytically show that high distortion of FW attacks is equivalent to small gradient variation along the attack path. It is then experimentally demonstrated on various deep neural network architectures that l-inf attacks against robust models achieve near maximal l-2 distortion, while standard networks have lower distortion. Furthermore, it is experimentally shown that catastrophic overfitting is strongly correlated with low distortion of FW attacks. This mathematical transparency differentiates FW from the more popular Projected Gradient Descent (PGD) optimization. To demonstrate the utility of our theoretical framework we develop FW-AT-Adapt, a novel adversarial training algorithm which uses a simple distortion measure to adapt the number of attack steps during training to increase efficiency without compromising robustness. FW-AT-Adapt provides training time on par with single-step fast AT methods and improves closing the gap between fast AT methods and multi-step PGD-AT with minimal loss in adversarial accuracy in white-box and black-box settings.
****

Certified Patch Robustness via Smoothed Vision Transformers
Hadi Salman, Saachi Jain, Eric Wong, Aleksander Madry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15137-15147
Certified patch defenses can guarantee robustness of an image classifier to arbitrary changes within a bounded contiguous region. But, currently, this robustness comes at a cost of degraded standard accuracies and slower inference times. We demonstrate how using vision transformers enables significantly better certified patch robustness that is also more computationally efficient and does not incur a substantial drop in standard accuracy. These improvements stem from the inherent ability of the vision transformer to gracefully handle largely masked images.
****

Look Back and Forth: Video Super-Resolution With Explicit Temporal Difference Modeling
Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Visi

on and Pattern Recognition (CVPR), 2022, pp. 17411-17420
Temporal modeling is crucial for video super-resolution. Most of the video super-resolution methods adopt the optical flow or deformable convolution for explicitly motion compensation. However, such temporal modeling techniques increase the model complexity and might fail in case of occlusion or complex motion, resulting in serious distortion and artifacts. In this paper, we propose to explore the role of explicit temporal difference modeling in both LR and HR space. Instead of directly feeding consecutive frames into a VSR model, we propose to compute the temporal difference between frames and divide those pixels into two subsets according to the level of difference. They are separately processed with two branches of different receptive fields in order to better extract complementary information. To further enhance the super-resolution result, not only spatial residual features are extracted, but the difference between consecutive frames in high-frequency domain is also computed. It allows the model to exploit intermediate SR results in both future and past to refine the current SR output. The difference at different time steps could be cached such that information from further distance in time could be propagated to the current frame for refinement. Experiments on several video super-resolution benchmark datasets demonstrate the effectiveness of the proposed method and its favorable performance against state-of-the-art methods.
****
UCC: Uncertainty Guided Cross-Head Co-Training for Semi-Supervised Semantic Segmentation
Jiashuo Fan, Bin Gao, Huan Jin, Lihui Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9947-9956
Deep neural networks (DNNs) have witnessed great successes in semantic segmentation, which requires a large number of labeled data for training. We present a novel learning framework called Uncertainty guided Cross-head Co-training (UCC) for semi-supervised semantic segmentation. Our framework introduces weak and strong augmentations within a shared encoder to achieve co-training, which naturally combines the benefits of consistency and self-training. Every segmentation head interacts with its peers and, the weak augmentation result is used for supervising the strong. The consistency training samples' diversity can be boosted by Dynamic Cross-Set Copy-Paste (DCSCP), which also alleviates the distribution mismatch and class imbalance problems. Moreover, our proposed Uncertainty Guided Re-weight Module (UGRM) enhances the self-training pseudo labels by suppressing the effect of the low-quality pseudo labels from its peer via modeling uncertainty. Extensive experiments on Cityscapes and PASCAL VOC 2012 demonstrate the effectiveness of our UCC, our approach significantly outperforms other state-of-the-art semi-supervised semantic segmentation methods. It achieves 77.17%, 76.49% mIoU on Cityscapes and PASCAL VOC 2012 datasets respectively under 1/16 protocols, which are +10.1%, +7.91% better than the supervised baseline.
****
HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture
Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhöfer, Jessica Hodgins, Christoph Lassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6143-6154
Capturing and rendering life-like hair is particularly challenging due to its fine geometric structure, complex physical interaction and the non-trivial visual appearance that must be captured. Yet, it is a critical component to create believable avatars. In this paper, we address the aforementioned problems: 1) we use a novel, volumetric hair representation that is composed of thousands of primitives. Each primitive can be rendered efficiently, yet realistically, by building on the latest advances in neural rendering. 2) To have a reliable control signal, we present a novel way of tracking hair on strand level. To keep the computational effort manageable, we use guide hairs and classic techniques to expand those into a dense head of hair. 3) To better enforce temporal consistency and generalization ability of our model, we further optimize the 3D scene flow of our representation with multiview optical flow, using volumetric raymarching. Our meth

od can not only create realistic renders of recorded multi-view sequences, but a lso create renderings for new hair configurations by providing new control signa ls. We compare our method with existing work on viewpoint synthesis and drivable animation and achieve state-of-the-art results. https://ziyanw1.github.io/hvh/
****

RADU: Ray-Aligned Depth Update Convolutions for ToF Data Denoising
Michael Schelling, Pedro Hermosilla, Timo Ropinski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 671-680
Time-of-Flight (ToF) cameras are subject to high levels of noise and distortions due to Multi-Path-Interference (MPI). While recent research showed that 2D neur al networks are able to outperform previous traditional State-of-the-Art (SOTA) methods on correcting ToF-Data, little research on learning-based approaches has been done to make direct use of the 3D information present in depth images. In this paper, we propose an iterative correcting approach operating in 3D space, t hat is designed to learn on 2.5D data by enabling 3D point convolutions to corre ct the points' positions along the view direction. As labeled real world data is scarce for this task, we further train our network with a self-training approac h on unlabeled real world data to account for real world statistics. We demonstr ate that our method is able to outperform SOTA methods on several datasets, incl uding two real world datasets and a new large-scale synthetic data set introduce d in this paper.
****

Rethinking Visual Geo-Localization for Large-Scale Applications
Gabriele Berton, Carlo Masone, Barbara Caputo; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4878-4888
Visual Geo-localization (VG) is the task of estimating the position where a give n photo was taken by comparing it with a large database of images of known locat ions. To investigate how existing techniques would perform on a real-world city-wide VG application, we build San Francisco eXtra Large, a new dataset covering a whole city and providing a wide range of challenging cases, with a size 30x bi gger than the previous largest dataset for visual geo-localization. We find that current methods fail to scale to such large datasets, therefore we design a new highly scalable training technique, called CosPlace, which casts the training a s a classification problem avoiding the expensive mining needed by the commonly used contrastive learning. We achieve state-of-the-art performance on a wide ran ge of datasets, and find that CosPlace is robust to heavy domain changes. Moreov er, we show that, compared to previous state of the art, CosPlace requires rough ly 80% less GPU memory at train time and achieves better results with 8x smaller descriptors, paving the way for city-wide real-world visual geo-localization. D ataset, code and trained models are available for research purposes at https://g ithub.com/gmberton/CosPlace.
****

Learning Based Multi-Modality Image and Video Compression
Guo Lu, Tianxiong Zhong, Jing Geng, Qiang Hu, Dong Xu; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6083-6092
Multi-modality (i.e., multi-sensor) data is widely used in various vision tasks for more accurate or robust perception. However, the increased data modalities b ring new challenges for data storage and transmission. The existing data compres sion approaches usually adopt individual codecs for each modality without consid ering the correlation between different modalities. This work proposes a multi-m odality compression framework for infrared and visible image pairs by exploiting the cross-modality redundancy. Specifically, given the image in the reference m odality (e.g., the infrared image), we use the channel-wise alignment module to produce the aligned features based on the affine transform. Then the aligned fea ture is used as the context information for compressing the image in the current modality (e.g., the visible image), and the corresponding affine coefficients a re losslessly compressed at negligible cost. Furthermore, we introduce the Trans former-based spatial alignment module to exploit the correlation between the int ermediate features in the decoding procedures for different modalities. Our fram

ework is very flexible and easily extended for multi-modality video compression. Experimental results show our proposed framework outperforms the traditional and learning-based single modality compression methods on the FLIR and KAIST datasets.

****

A Stitch in Time Saves Nine: A Train-Time Regularizing Loss for Improved Neural Network Calibration

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, Chetan Arora; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16081-16090

Deep Neural Networks (DNNs) are known to make overconfident mistakes, which makes their use problematic in safety-critical applications. State-of-the-art (SOTA) calibration techniques improve on the confidence of predicted labels alone, and leave the confidence of non-max classes (e.g. top-2, top-5) uncalibrated. Such calibration is not suitable for label refinement using post-processing. Further, most SOTA techniques learn a few hyper-parameters post-hoc, leaving out the scope for image, or pixel specific calibration. This makes them unsuitable for calibration under domain shift, or for dense prediction tasks like semantic segmentation. In this paper, we argue for intervening at the train time itself, so as to directly produce calibrated DNN models. We propose a novel auxiliary loss function: Multi-class Difference in Confidence and Accuracy (MDCA), to achieve the same. MDCA can be used in conjunction with other application/task specific loss functions. We show that training with MDCA leads to better calibrated models in terms of Expected Calibration Error (ECE), and Static Calibration Error (SCE) on image classification, and segmentation tasks. We report ECE(SCE) score of 0.72 (1.60) on the CIFAR100 dataset, in comparison to 1.90 (1.71) by the SOTA. Under domain shift, a ResNet-18 model trained on PACS dataset using MDCA gives a average ECE(SCE) score of 19.7 (9.7) across all domains, compared to 24.2 (11.8) by the SOTA. For segmentation task, we report a 2x reduction in calibration error on PASCAL-VOC dataset in comparison to Focal Loss. Finally, MDCA training improves calibration even on imbalanced data, and for natural language classification tasks.

****

The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy

Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, Zhangyang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12020-12030

Vision transformers (ViTs) have gained increasing popularity as they are commonly believed to own higher modeling capacity and representation flexibility, than traditional convolutional networks. However, it is questionable whether such potential has been fully unleashed in practice, as the learned ViTs often suffer from over-smoothening, yielding likely redundant models. Recent works made preliminary attempts to identify and alleviate such redundancy, e.g., via regularizing embedding similarity or re-injecting convolution-like structures. However, a "head-to-toe assessment" regarding the extent of redundancy in ViTs, and how much we could gain by thoroughly mitigating such, has been absent for this field. This paper, for the first time, systematically studies the ubiquitous existence of redundancy at all three levels: patch embedding, attention map, and weight space. In view of them, we advocate a principle of diversity for training ViTs, by presenting corresponding regularizers that encourage the representation diversity and coverage at each of those levels, that enabling capturing more discriminative information. Extensive experiments on ImageNet with a number of ViT backbones validate the effectiveness of our proposals, largely eliminating the observed ViT redundancy and significantly boosting the model generalization. For example, our diversified DeiT obtains 0.70% 1.76% accuracy boosts on ImageNet with highly reduced similarity. Our codes are fully available in https://github.com/VITA-Group/Diverse-ViT.

****

Deep Image-Based Illumination Harmonization

Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, Chu nxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18542-18551

Integrating a foreground object into a background scenewith illumination harmoni zation is an important but chal-lenging task in computer vision and augmented re ality community. Existing methods mainly focus on foreground andbackground appea rance consistency or the foreground object shadow generation, which rarely consi der global appearance and illumination harmonization. In this paper,we formulate seamless illumination harmonization as anillumination exchange and aggregation problem. Specifi-cally, we firstly apply a physically-based rendering methodto c onstruct a large-scale, high-quality dataset (named IH)for our task, which conta ins various types of foreground ob-jects and background scenes with different li ghting conditions. Then, we propose a deep image-based illuminationharmonization GAN framework named DIH-GAN, whichmakes full use of a multi-scale attention mec hanism and illumination exchange strategy to directly infer mapping rela-tionshi p between the inserted foreground object and the corresponding background scene. Meanwhile, we also use adversarial learning strategy to further refine the illu minationharmonization result. Our method can not only achieve har-monious appear ance and illumination for the foregroundobject but also can generate compelling shadow cast bythe foreground object. Comprehensive experiments on bothour IH dat aset and real-world images show that our pro-posed DIH-GAN provides a practical and effective solutionfor image-based object illumination harmonization editing, and validate the superiority of our method against state-of-the-art methods.

****

ViM: Out-of-Distribution With Virtual-Logit Matching

Haoqi Wang, Zhizhong Li, Litong Feng, Wayne Zhang; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4921-4930

Most of the existing Out-Of-Distribution (OOD) detection algorithms depend on si ngle input source: the feature, the logit, or the softmax probability. However, the immense diversity of the OOD examples makes such methods fragile. There are OOD samples that are easy to identify in the feature space while hard to disting uish in the logit space and vice versa. Motivated by this observation, we propos e a novel OOD scoring method named Virtual-logit Matching (ViM), which combines the class-agnostic score from feature space and the In-Distribution (ID) class-d ependent logits. Specifically, an additional logit representing the virtual OOD class is generated from the residual of the feature against the principal space, and then matched with the original logits by a constant scaling. The probabilit y of this virtual logit after softmax is the indicator of OOD-ness. To facilitat e the evaluation of large-scale OOD detection in academia, we create a new OOD d ataset for ImageNet-1K, which is human-annotated and is 8.8x the size of existin g datasets. We conducted extensive experiments, including CNNs and vision transf ormers, to demonstrate the effectiveness of the proposed ViM score. In particula r, using the BiT-S model, our method gets an average AUROC 90.91% on four diffic ult OOD benchmarks, which is 4% ahead of the best baseline. Code and dataset are available at https://github.com/haoqiwang/vim.

****

Active Learning by Feature Mixing

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza (Reza) Haffari, Anton van den Hengel, Javen Qinfeng Shi; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2022, pp. 12237-12246

The promise of active learning (AL) is to reduce labelling costs by selecting th e most valuable examples to annotate from a pool of unlabelled data. Identifying these examples is especially challenging with high-dimensional data (e. g. imag es, videos) and in low-data regimes. In this paper, we propose a novel method fo r batch AL called ALFA-Mix. We identify unlabelled instances with sufficiently-d istinct features by seeking inconsistencies in predictions resulting from interv entions on their representations. We construct interpolations between representa tions of labelled and unlabelled instances then examine the predicted labels. We show that inconsistencies in these predictions help discovering features that t he model is unable to recognise in the unlabelled instances. We derive an effici

ent implementation based on a closed-form solution to the optimal interpolation causing changes in predictions. Our method outperforms all recent AL approaches in 30 different settings on 12 benchmarks of images, videos, and non-visual data. The improvements are especially significant in low-data regimes and on self-trained vision transformers, where ALFA-Mix outperforms the state-of-the-art in 59% and 43% of the experiments respectively.

****

Towards Accurate Facial Landmark Detection via Cascaded Transformers

Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, Jae-Joon Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4176-4185

Accurate facial landmarks are essential prerequisites for many tasks related to human faces. In this paper, an accurate facial landmark detector is proposed based on cascaded transformers. We formulate facial landmark detection as a coordinate regression task such that the model can be trained end-to-end. With self-attention in transformers, our model can inherently exploit the structured relationships between landmarks, which would benefit landmark detection under challenging conditions such as large pose and occlusion. During cascaded refinement, our model is able to extract the most relevant image features around the target landmark for coordinate prediction, based on deformable attention mechanism, thus bringing more accurate alignment. In addition, we propose a novel decoder that refines image features and landmark positions simultaneously. With few parameter increasing, the detection performance improves further. Our model achieves new state-of- the-art performance on several standard facial landmark detection benchmarks, and shows good generalization ability in cross-dataset evaluation.

****

Class-Aware Contrastive Semi-Supervised Learning

Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, Long Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14421-14430

Pseudo-label-based semi-supervised learning (SSL) has achieved great success on raw data utilization. However, its training procedure suffers from confirmation bias due to the noise contained in self-generated artificial labels. Moreover, the model's judgment becomes noisier in real-world applications with extensive out-of-distribution data. To address this issue, we propose a general method named Class-aware Contrastive Semi-Supervised Learning (CCSSL), which is a drop-in helper to improve the pseudo-label quality and enhance the model's robustness in the real-world setting. Rather than treating real-world data as a union set, our method separately handles reliable in-distribution data with class-wise clustering for blending into downstream tasks and noisy out-of-distribution data with image-wise contrastive for better generalization. Furthermore, by applying target re-weighting, we successfully emphasize clean label learning and simultaneously reduce noisy label learning. Despite its simplicity, our proposed CCSSL has significant performance improvements over the state-of-the-art SSL methods on the standard datasets CIFAR100 and STL10. On the real-world dataset Semi-iNat 2021, we improve FixMatch by 9.80% and CoMatch by 3.18%. Code is available https://github.com/TencentYoutuResearch/Classification-SemiCLS.

****

Long-Term Visual Map Sparsification With Heterogeneous GNN

Ming-Fang Chang, Yipu Zhao, Rajvi Shah, Jakob J. Engel, Michael Kaess, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2406-2415

We address the problem of map sparsification for longterm visual localization. A commonly employed assumption in map sparsification is that the pre-build map and the later capture localization query are consistent. However, this assumption can be easily violated in the dynamic world. Additionally, the map size grows as new data accumulate through time, causing large data overhead in the long term. In this paper, we aim to overcome the environmental changes and reduce the map size at the same time by selecting points that are valuable to future localization. Inspired by the recent progress in Graph Neural Network (GNN), we propose th

e first work that models SfM maps as heterogeneous graphs and predicts 3D point importance scores with a GNN, which enables us to directly exploit the rich info rmation in the SfM map graph. Two novel supervisions are proposed: 1) a data-fit ting term for selecting valuable points to future localization based on training queries; 2) a K-Cover term for selecting sparse points with full-map coverage. In the experiments on a long-term dataset with environmental changes, our method selected map points on stable and widely visible structures and outperformed ba selines in localization performance. This work novelly connects SfM maps with th e abundant modern GNN techniques and opens a new research avenue forward.
****

Debiased Learning From Naturally Imbalanced Pseudo-Labels
Xudong Wang, Zhirong Wu, Long Lian, Stella X. Yu; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14647-1465 7
This work studies the bias issue of pseudo-labeling, a natural phenomenon that w idely occurs but often overlooked by prior research. Pseudo-labels are generated when a classifier trained on source data is transferred to unlabeled target dat a. We observe heavy long-tailed pseudo-labels when a semi-supervised learning mo del FixMatch predicts labels on the unlabeled set even though the unlabeled data is curated to be balanced. Without intervention, the training model inherits th e bias from the pseudo-labels and end up being sub-optimal. To eliminate the mod el bias, we propose a simple yet effective method DebiasMatch, comprising of an adaptive debiasing module and an adaptive marginal loss. The strength of debiasi ng and the size of margins can be automatically adjusted by making use of an onl ine updated queue. Benchmarked on ImageNet-1K, DebiasMatch significantly outperf orms previous state-of-the-arts by more than 26% and 10.5% on semi-supervised le arning (0.2% annotated data) and zero-shot learning tasks respectively.
****

RNNPose: Recurrent 6-DoF Object Pose Refinement With Robust Correspondence Field Estimation and Pose Optimization
Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 14880-14890
6-DoF object pose estimation from a monocular image is challenging, and a post-r efinement procedure is generally needed for high-precision estimation. In this p aper, we propose a framework based on a recurrent neural network (RNN) for objec t pose refinement, which is robust to erroneous initial poses and occlusions. Du ring the recurrent iterations, object pose refinement is formulated as a non-lin ear least squares problem based on the estimated correspondence field (between a rendered image and the observed image). The problem is then solved by a differe ntiable Levenberg-Marquardt (LM) algorithm enabling end-to-end training. The cor respondence field estimation and pose refinement are conducted alternatively in each iteration to recover the object poses. Furthermore, to improve the robustne ss to occlusion, we introduce a consistency-check mechanism based on the learned descriptors of the 3D model and observed 2D images, which downweights the unrel iable correspondences during pose optimization. Extensive experiments on LINEMOD , Occlusion-LINEMOD, and YCB-Video datasets validate the effectiveness of our me thod and demonstrate state-of-the-art performance.
****

Ditto: Building Digital Twins of Articulated Objects From Interaction
Zhenyu Jiang, Cheng-Chun Hsu, Yuke Zhu; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5616-5626
Digitizing physical objects into the virtual world has the potential to unlock n ew research and applications in embodied AI and mixed reality. This work focuses on recreating interactive digital twins of real-world articulated objects, whic h can be directly imported into virtual environments. We introduce Ditto to lear n articulation model estimation and 3D geometry reconstruction of an articulated object through interactive perception. Given a pair of visual observations of a n articulated object before and after interaction, Ditto reconstructs part-level geometry and estimates the articulation model of the object. We employ implicit

neural representations for joint geometry and articulation modeling. Our experiments show that Ditto effectively builds digital twins of articulated objects in a category-agnostic way. We also apply Ditto to real-world objects and deploy the recreated digital twins in physical simulation. Code and additional results are available at https://ut-austin-rpl.github.io/Ditto/

****

Dual-AI: Dual-Path Actor Interaction Learning for Group Activity Recognition

Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2990-2999

Learning spatial-temporal relation among multiple actors is crucial for group activity recognition. Different group activities often show the diversified interactions between actors in the video. Hence, it is often difficult to model complex group activities from a single view of spatial-temporal actor evolution. To tackle this problem, we propose a distinct Dual-path Actor Interaction (Dual-AI) framework, which flexibly arranges spatial and temporal transformers in two complementary orders, enhancing actor relations by integrating merits from different spatio-temporal paths. Moreover, we introduce a novel Multi-scale Actor Contrastive Loss (MAC-Loss) between two interactive paths of Dual-AI. Via self-supervised actor consistency in both frame and video levels, MAC-Loss can effectively distinguish individual actor representations to reduce action confusion among different actors. Consequently, our Dual-AI can boost group activity recognition by fusing such discriminative features of different actors. To evaluate the proposed approach, we conduct extensive experiments on the widely used benchmarks, including Volleyball, Collective Activity, and NBA datasets. The proposed Dual-AI achieves state-of-the-art performance on all these datasets. It is worth noting the proposed Dual-AI with 50% training data outperforms a number of recent approaches with 100% training data. This confirms the generalization power of Dual-AI for group activity recognition, even under the challenging scenarios of limited supervision.

****

Harmony: A Generic Unsupervised Approach for Disentangling Semantic Content From Parameterized Transformations

Mostofa Rafid Uddin, Gregory Howe, Xiangrui Zeng, Min Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20646-20655

In many real-life image analysis applications, particularly in biomedical research domains, the objects of interest undergo multiple transformations that alters their visual properties while keeping the semantic content unchanged. Disentangling images into semantic content factors and transformations can provide significant benefits into many domain-specific image analysis tasks. To this end, we propose a generic unsupervised framework, Harmony, that simultaneously and explicitly disentangles semantic content from multiple parameterized transformations. Harmony leverages a simple cross-contrastive learning framework with multiple explicitly parameterized latent representations to disentangle content from transformations. To demonstrate the efficacy of Harmony, we apply it to disentangle image semantic content from several parameterized transformations (rotation, translation, scaling, and contrast). Harmony achieves significantly improved disentanglement over the baseline models on several image datasets of diverse domains. With such disentanglement, Harmony is demonstrated to incentivize bioimage analysis research by modeling structural heterogeneity of macromolecules from cryo-ET images and learning transformation-invariant representations of protein particles from single-particle cryo-EM images. Harmony also performs very well in disentangling content from 3D transformations and can perform coarse and fast alignment of 3D cryo-ET subtomograms. Therefore, Harmony is generalizable to many other imaging domains and can potentially be extended to domains beyond imaging as well.

****

Talking Face Generation With Multilingual TTS

Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseo

ng Lee, Dongho Choi, Kang-wook Kim; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 21425-21430
Recent studies in talking face generation have focused on building a model that
can generalize from any source speech to any target identity. A number of works
have already claimed this functionality and have added that their models will al
so generalize to any language. However, we show, using languages from different
language families, that these models do not translate well when the training lan
guage and the testing language are sufficiently different. We reduce the scope o
f the problem to building a languagerobust talking face generation system on see
n identities, i.e., the target identity is the same as the training identity. In
 this work, we introduce a talking face generation system that generalizes to di
fferent languages. We evaluate the efficacy of our system using a multilingual t
ext-to-speech system. We present the joint text-to-speech system and the talking
 face generation system as a neural dubber system. Our demo is available at http
s://bit.ly/ml-face-generation-cvpr22-demo. Also, our screencast is uploaded at h
ttps://youtu.be/F6h0s0M4vBI.
****

A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by M
ultiple Dance Genres
Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, Sanghoon Lee; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 3490-3500
When coming up with phrases of movement, choreographers all have their habits as
 they are used to their skilled dance genres. Therefore, they tend to return cer
tain patterns of the dance genres that they are familiar with. What if artificia
l intelligence could be used to help choreographers blend dance genres by sugges
ting various dances, and one that matches their choreographic style? Numerous ta
sk-specific variants of autoregressive networks have been developed for dance ge
neration. Yet, a serious limitation remains that all existing algorithms can ret
urn repeated patterns for a given initial pose sequence, which may be inferior.
To mitigate this issue, we propose MNET, a novel and scalable approach that can
perform music-conditioned pluralistic dance generation synthesized by multiple d
ance genres using only a single model. Here, we learn a dance-genre aware latent
 representation by training a conditional generative adversarial network leverag
ing Transformer architecture. We conduct extensive experiments on AIST++ along w
ith user studies. Compared to the state-of-the-art methods, our method synthesiz
es plausible and diverse outputs according to multiple dance genres as well as g
enerates outperforming dance sequences qualitatively and quantitatively.
****

Kernelized Few-Shot Object Detection With Efficient Integral Aggregation
Shan Zhang, Lei Wang, Naila Murray, Piotr Koniusz; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19207-192
16
We design a Kernelized Few-shot Object Detector by leveraging kernelized matrice
s computed over multiple proposal regions, which yield expressive non-linear rep
resentations whose model complexity is learned on the fly. Our pipeline contains
 several modules. An Encoding Network encodes support and query images. Our Kern
elized Autocorrelation unit forms the linear, polynomial and RBF kernelized repr
esentations from features extracted within support regions of support images. Th
ese features are then cross-correlated against features of a query image to obta
in attention weights, and generate query proposal regions via an Attention Regio
n Proposal Net. As the query proposal regions are many, each described by the li
near, polynomial and RBF kernelized matrices, their formation is costly but that
 cost is reduced by our proposed Integral Region-of-Interest Aggregation unit. F
inally, the Multi-head Relation Net combines all kernelized (second-order) repre
sentations with the first-order feature maps to learn support-query class relati
ons and locations. We outperform the state of the art on novel classes by 3.8%,
5.4% and 5.7% mAP on PASCAL VOC 2007, FSOD, and COCO.
****

Transformer Based Line Segment Classifier With Image Context for Real-Time Vanis

hing Point Detection in Manhattan World
Xin Tong, Xianghua Ying, Yongjie Shi, Ruibin Wang, Jinfa Yang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 6093-6102
Previous works on vanishing point detection usually use geometric prior for line segment clustering. We find that image context can also contribute to accurate line classification. Based on this observation, we propose to classify line segm ents into three groups according to three unknown-but-sought vanishing points wi th Manhattan world assumption, using both geometric information and image contex t in this work. To achieve this goal, we propose a novel Transformer based Line segment Classifier (TLC) that can group line segments in images and estimate the corresponding vanishing points. In TLC, we design a line segment descriptor to represent line segments using their positions, directions and local image contex ts. Transformer based feature fusion module is used to capture global features f rom all line segments, which is proved to improve the classification performance significantly in our experiments. By using a network to score line segments for outlier rejection, vanishing points can be got by Singular Value Decomposition (SVD) from the classified lines. The proposed method runs at 25 fps on one NVIDI A 2080Ti card for vanishing point detection. Experimental results on synthetic a nd real-world datasets demonstrate that our method is superior to other state-of -the-art methods on the balance between accuracy and efficiency, while keeping s tronger generalization capability when trained and evaluated on different datase ts.
****

Self-Sustaining Representation Expansion for Non-Exemplar Class-Incremental Lear ning
Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, Zheng-Jun Zha; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9296-9305
Non-exemplar class-incremental learning is to recognize both the old and new cla sses when old class samples cannot be saved. It is a challenging task since repr esentation optimization and feature retention can only be achieved under supervi sion from new classes. To address this problem, we propose a novel self-sustaini ng representation expansion scheme. Our scheme consists of a structure reorganiz ation strategy that fuses main-branch expansion and side-branch updating to main tain the old features, and a main-branch distillation scheme to transfer the inv ariant knowledge. Furthermore, a prototype selection mechanism is proposed to en hance the discrimination between the old and new classes by selectively incorpor ating new samples into the distillation process. Extensive experiments on three benchmarks demonstrate significant incremental performance, outperforming the st ate-of-the-art methods by a margin of 3% , 3% and 6% , respectively.
****

Adaptive Early-Learning Correction for Segmentation From Noisy Annotations
Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, Carlos Fernandez-Granda; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 2606-2616
Deep learning in the presence of noisy annotations has been studied extensively in classification, but much less in segmentation tasks. In this work, we study t he learning dynamics of deep segmentation networks trained on inaccurately-annot ated data. We discover a phenomenon that has been previously reported in the con text of classification: the networks tend to first fit the clean pixel-level lab els during an "early-learning" phase, before eventually memorizing the false ann otations. However, in contrast to classification, memorization in segmentation d oes not arise simultaneously for all semantic categories. Inspired by these find ings, we propose a new method for segmentation from noisy annotations with two k ey elements. First, we detect the beginning of the memorization phase separately for each category during training. This allows us to adaptively correct the noi sy annotations in order to exploit early learning. Second, we incorporate a regu larization term that enforces consistency across scales to boost robustness agai nst annotation noise. Our method outperforms standard approaches on a medical-im

aging segmentation task where noises are synthesized to mimic human annotation e
rrors. It also provides robustness to realistic noisy annotations present in wea
kly-supervised semantic segmentation, achieving state-of-the-art results on PASC
AL VOC 2012. Code is available at https://github.com/Kangningthu/ADELE
****

Cross-Domain Correlation Distillation for Unsupervised Domain Adaptation in Nigh
ttime Semantic Segmentation
Huan Gao, Jichang Guo, Guoli Wang, Qian Zhang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9913-9923
The performance of nighttime semantic segmentation is restricted by the poor ill
umination and a lack of pixel-wise annotation, which severely limit its applicat
ion in autonomous driving. Existing works, e.g., using the twilight as the inter
mediate target domain to perform the adaptation from daytime to nighttime, may f
ail to cope with the inherent difference between datasets caused by the camera e
quipment and the urban style. Faced with these two types of domain shifts, i.e.,
 the illumination and the inherent difference of the datasets, we propose a nove
l domain adaptation framework via cross-domain correlation distillation, called
CCDistill. The invariance of illumination or inherent difference between two ima
ges is fully explored so as to make up for the lack of labels for nighttime imag
es. Specifically, we extract the content and style knowledge contained in featur
es, calculate the degree of inherent or illumination difference between two imag
es. The domain adaptation is achieved using the invariance of the same kind of d
ifference. Extensive experiments on Dark Zurich and ACDC demonstrate that CCDist
ill achieves the state-of-the-art performance for nighttime semantic segmentatio
n. Notably, our method is a one-stage domain adaptation network which can avoid
affecting the inference time. Our implementation is available at https://github.
com/ghuan99/CCDistill.
****

Context-Aware Video Reconstruction for Rolling Shutter Cameras
Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, Mingyi He; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17572
-17582
With the ubiquity of rolling shutter (RS) cameras, it is becoming increasingly a
ttractive to recover the latent global shutter (GS) video from two consecutive R
S frames, which also places a higher demand on realism. Existing solutions, usin
g deep neural networks or optimization, achieve promising performance. However,
these methods generate intermediate GS frames through image warping based on the
 RS model, which inevitably result in black holes and noticeable motion artifact
s. In this paper, we alleviate these issues by proposing a context-aware GS vide
o reconstruction architecture. It facilitates the advantages such as occlusion r
easoning, motion compensation, and temporal abstraction. Specifically, we first
estimate the bilateral motion field so that the pixels of the two RS frames are
warped to a common GS frame accordingly. Then, a refinement scheme is proposed t
o guide the GS frame synthesis along with bilateral occlusion masks to produce h
igh-fidelity GS video frames at arbitrary times. Furthermore, we derive an appro
ximated bilateral motion field model, which can serve as an alternative to provi
de a simple but effective GS frame initialization for related tasks. Experiments
 on synthetic and real data show that our approach achieves superior performance
 over state-of-the-art methods in terms of objective metrics and subjective visu
al quality.
****

Towards Efficient Data Free Black-Box Adversarial Attack
Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, Chao Wu; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 15115-15125
Classic black-box adversarial attacks can take advantage of transferable adversa
rial examples generated by a similar substitute model to successfully fool the t
arget model. However, these substitute models need to be trained by target model
s' training data, which is hard to acquire due to privacy or transmission reason
s. Recognizing the limited availability of real data for adversarial queries, re

cent works proposed to train substitute models in a data-free black-box scenario . However, their generative adversarial networks (GANs) based framework suffers from the convergence failure and the model collapse, resulting in low efficiency . In this paper, by rethinking the collaborative relationship between the genera tor and the substitute model, we design a novel black-box attack framework. The proposed method can efficiently imitate the target model through a small number of queries and achieve high attack success rate. The comprehensive experiments o ver six datasets demonstrate the effectiveness of our method against the state-o f-the-art attacks. Especially, we conduct both label-only and probability-only a ttacks on the Microsoft Azure online model, and achieve a 100% attack success ra te with only 0.46% query budget of the SOTA method [??].
****

Robust Contrastive Learning Against Noisy Views
Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio To rralba, Stefanie Jegelka, Yale Song; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2022, pp. 16670-16681
Contrastive learning relies on an assumption that positive pairs contain related views that share certain underlying information about an instance, e.g., patche s of an image or co-occurring multimodal signals of a video. What if this assump tion is violated? The literature suggests that contrastive learning produces sub optimal representations in the presence of noisy views, e.g., false positive pai rs with no apparent shared information. In this work, we propose a new contrasti ve loss function that is robust against noisy views. We provide rigorous theoret ical justifications by showing connections to robust symmetric losses for noisy binary classification and by establishing a new contrastive bound for mutual inf ormation maximization based on the Wasserstein distance measure. The proposed lo ss is completely modality-agnostic and a simple drop-in replacement for the Info NCE loss, which makes it easy to apply to existing contrastive frameworks. We sh ow that our approach provides consistent improvements over the state-of-the-art on image, video, and graph contrastive learning benchmarks that exhibit a variet y of real-world noise patterns.
****

More Than Words: In-the-Wild Visually-Driven Prosody for Text-to-Speech
Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, Tal Remez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10587-10597
In this paper we present VDTTS, a Visually-Driven Text-to-Speech model. Motivate d by dubbing, VDTTS takes advantage of video frames as an additional input along side text, and generates speech that matches the video signal. We demonstrate ho w this allows VDTTS to, unlike plain TTS models, generate speech that not only h as prosodic variations like natural pauses and pitch, but is also synchronized t o the input video. Experimentally, we show our model produces well-synchronized outputs, approaching the video-speech synchronization quality of the ground-trut h, on several challenging benchmarks including "in-the-wild" content from VoxCel eb2. Supplementary demo videos demonstrating video-speech synchronization, robus tness to speaker ID swapping, and prosody, presented at the project page.
****

Cross-Modal Perceptionist: Can Face Geometry Be Gleaned From Voices?
Cho-Ying Wu, Chin-Cheng Hsu, Ulrich Neumann; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10452-10461
This work digs into a root question in human perception: can face geometry be gl eaned from one's voices? Previous works that study this question only adopt deve lopments in image synthesis and convert voices into face images to show correlat ions, but working on the image domain unavoidably involves predicting attributes that voices cannot hint, including facial textures, hairstyles, and backgrounds . We instead investigate the ability to reconstruct 3D faces to concentrate on o nly geometry, which is much more physiologically grounded. We propose our analys is framework, Cross-Modal Perceptionist, under both supervised and unsupervised learning. First, we construct a dataset, Voxceleb-3D, which extends Voxceleb and includes paired voices and face meshes, making supervised learning possible. Se

cond, we use a knowledge distillation mechanism to study whether face geometry c
an still be gleaned from voices without paired voices and 3D face data under lim
ited availability of 3D face scans. We break down the core question into four pa
rts and perform visual and numerical analyses as responses to the core question.
 Our findings echo those in physiology and neuroscience about the correlation be
tween voices and facial structures. The work provides future human-centric cross
-modal learning with explainable foundations. See our project page.
****

On Generalizing Beyond Domains in Cross-Domain Continual Learning
Christian Simon, Masoud Faraki, Yi-Hsuan Tsai, Xiang Yu, Samuel Schulter, Yumin
Suh, Mehrtash Harandi, Manmohan Chandraker; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9265-9274
In the real world, humans have the ability to accumulate new knowledge in any co
nditions. However, deeplearning suffers from the phenomenon so-called catastroph
ic forgetting of the previously observed knowledge after learning a new task. Ma
ny recent methods focus on preventing catastrophic forgetting under a typical as
sumption of thetrain and test data following a similar distribution. In thiswork
, we consider the more realistic scenario of continuallearning under domain shif
ts where the model is able to gen-eralize its inference to a an unseen domain. T
o this end, wepropose to make use of sample correlations of the learning tasks i
n the classifiers where the subsequent optimization isperformed over similarity
measures obtained in a similar fashion to the Mahalanobis distance computation.
In addition, we also propose an approach based on the exponential moving average
 of the parameters for better knowledge distillation, allowing a further adaptat
ion to the old model. We demonstrate in our experiments that, to a great extent,
 the past continual learning algorithms fail to handle the forgetting issue unde
r multiple distributions, while our proposed approach identifies the task under
domain shift where insome cases can boost up the performance up to 10% on the ch
allenging datasets e.g., DomainNet and OfficeHome.
****

RSTT: Real-Time Spatial Temporal Transformer for Space-Time Video Super-Resoluti
on
Zhicheng Geng, Luming Liang, Tianyu Ding, Ilya Zharkov; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1744
1-17451
Space-time video super-resolution (STVSR) is the task of interpolating videos wi
th both Low Frame Rate (LFR) and Low Resolution (LR) to produce High-Frame-Rate
(HFR) and also High-Resolution (HR) counterparts. The existing methods based on
Convolutional Neural Network (CNN) succeed in achieving visually satisfied resul
ts while suffer from slow inference speed due to their heavy architectures. We p
ropose to resolve this issue by using a spatial-temporal transformer that natura
lly incorporates the spatial and temporal super resolution modules into a single
 model. Unlike CNN-based methods, we do not explicitly use separated building bl
ocks for temporal interpolations and spatial super-resolutions; instead, we only
 use a single end-to-end transformer architecture. Specifically, a reusable dict
ionary is built by encoders based on the input LFR and LR frames, which is then
utilized in the decoder part to synthesize the HFR and HR frames. Compared with
the state-of-the-art TMNet, our network is 60% smaller (4.5M vs 12.3M parameters
) and 80% faster (26.2fps vs 14.3fps on 720 x 576 frames) without sacrificing mu
ch performance. The source code is available at https://github.com/llmpass/RSTT.
****

Learning Memory-Augmented Unidirectional Metrics for Cross-Modality Person Re-Id
entification
Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, Wenhui Li; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 19366-19375
This paper tackles the cross-modality person re-identification (re-ID) problem b
y suppressing the modality discrepancy. In cross-modality re-ID, the query and g
allery images are in different modalities. Given a training identity, the popula
r deep classification baseline shares the same proxy (i.e., a weight vector in t

he last classification layer) for two modalities. We find that it has considerab le tolerance for the modality gap, because the shared proxy acts as an intermedi ate relay between two modalities. In response, we propose a Memory-Augmented Uni directional Metric (MAUM) learning method consisting of two novel designs, i.e., unidirectional metrics, and memory-based augmentation. Specifically, MAUM first learns modality-specific proxies (MS-Proxies) independently under each modality . Afterward, MAUM uses the already-learned MS-Proxies as the static references f or pulling close the features in the counterpart modality. These two unidirectio nal metrics (IR image to RGB proxy and RGB image to IR proxy) jointly alleviate the relay effect and benefit cross-modality association. The cross-modality asso ciation is further enhanced by storing the MS-Proxies into memory banks to incre ase the reference diversity. Importantly, we show that MAUM improves cross-modal ity re-ID under the modality-balanced setting and gains extra robustness against the modality-imbalance problem. Extensive experiments on SYSU-MM01 and RegDB da tasets demonstrate the superiority of MAUM over the state-of-the-art. The code w ill be available.

****

A Closer Look at Few-Shot Image Generation

Yunqing Zhao, Henghui Ding, Houjing Huang, Ngai-Man Cheung; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9140-9150

Modern GANs excel at generating high-quality and diverse images. However, when t ransferring the pretrained GANs on small target data (e.g., 10-shot), the genera tor tends to replicate the training samples. Several methods have been proposed to address this few-shot image generation task, but there is a lack of effort to analyze them under a unified framework. As our first contribution, we propose a framework to analyze existing methods during the adaptation. Our analysis disco vers that while some methods have a disproportionate focus on diversity preservi ng which impedes quality improvement, all methods achieve similar quality after convergence. Therefore, the better methods are those that can slow down diversit y degradation. Furthermore, our analysis reveals that there is still plenty of r oom to further slow down diversity degradation. Informed by our analysis and to slow down diversity degradation of the target generator during adaptation, our s econd contribution proposes to apply mutual information (MI) maximization to ret ain the source domain's rich multi-level diversity information in the target dom ain generator. We propose to perform MI maximization by contrastive loss (CL), l everage the generator and discriminator as two feature encoders to extract diffe rent multi-level features for computing CL. We refer to our method as Dual Contr astiveLearning (DCL). Extensive experiments on several public datasets show that , while leading to a slower diversity-degrading generator during adaptation, our proposed DCL brings visually pleasant quality and state-of-the-art quantitative performance.

****

Depth-Supervised NeRF: Fewer Views and Faster Training for Free

Kangle Deng, Andrew Liu, Jun-Yan Zhu, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12882-12 891

A commonly observed failure mode of Neural Radiance Field (NeRF) is fitting inco rrect geometries when given an insufficient number of input views. One potential reason is that standard volumetric rendering does not enforce the constraint th at most of a scene's geometry consist of empty space and opaque surfaces. We for malize the above assumption through DS-NeRF (Depth-supervised Neural Radiance Fi elds), a loss for learning radiance fields that takes advantage of readily-avail able depth supervision. We leverage the fact that current NeRF pipelines require images with known camera poses that are typically estimated by running structur e-from-motion (SFM). Crucially, SFM also produces sparse 3D points that can be u sed as "free" depth supervision during training: we add a loss to encourage the distribution of a ray's terminating depth matches a given 3D keypoint, incorpora ting depth uncertainty. DS-NeRF can render better images given fewer training vi ews while training 2-3x faster. Further, we show that our loss is compatible wit

h other recently proposed NeRF methods, demonstrating that depth is a cheap and easily digestible supervisory signal. And finally, we find that DS-NeRF can supp ort other types of depth supervision such as scanned depth sensors and RGBD reco nstruction outputs.
****

Unsupervised Domain Generalization by Learning a Bridge Across Domains
Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, Dina Katabi, Kate Saenko, Rogerio S. Feris, Leonid Karlinsky; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5280-5290
The ability to generalize learned representations across significantly different visual domains, such as between real photos, clipart, paintings, and sketches, is a fundamental capacity of the human visual system. In this paper, different f rom most cross-domain works that utilize some (or full) source domain supervisio n, we approach a relatively new and very practical Unsupervised Domain Generaliz ation (UDG) setup of having no training supervision in neither source nor target domains. Our approach is based on self-supervised learning of a Bridge Across D omains (BrAD) - an auxiliary bridge domain accompanied by a set of semantics pre serving visual (image-to-image) mappings to BrAD from each of the training domai ns. The BrAD and mappings to it are learned jointly (end-to-end) with a contrast ive self-supervised representation model that semantically aligns each of the do mains to its BrAD-projection, and hence implicitly drives all the domains (seen or unseen) to semantically align to each other. In this work, we show how using an edge-regularized BrAD our approach achieves significant gains across multiple benchmarks and a range of tasks, including UDG, Few-shot UDA, and unsupervised generalization across multi-domain datasets (including generalization to unseen domains and classes).
****

Partial Class Activation Attention for Semantic Segmentation
Sun-Ao Liu, Hongtao Xie, Hai Xu, Yongdong Zhang, Qi Tian; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16 836-16845
Current attention-based methods for semantic segmentation mainly model pixel rel ation through pairwise affinity and coarse segmentation. For the first time, thi s paper explores modeling pixel relation via Class Activation Map (CAM). Beyond the previous CAM generated from image-level classification, we present Partial C AM, which subdivides the task into region-level prediction and achieves better l ocalization performance. In order to eliminate the intra-class inconsistency cau sed by the variances of local context, we further propose Partial Class Activati on Attention (PCAA) that simultaneously utilizes local and global class-level re presentations for attention calculation. Once obtained the partial CAM, PCAA col lects local class centers and computes pixel-to-class relation locally. Applying local-specific representations ensures reliable results under different local c ontexts. To guarantee global consistency, we gather global representations from all local class centers and conduct feature aggregation. Experimental results co nfirm that Partial CAM outperforms the previous two strategies as pixel relation . Notably, our method achieves state-of-the-art performance on several challengi ng benchmarks including Cityscapes, Pascal Context, and ADE20K. Code is availabl e at https://github.com/lsa1997/PCAA.
****

Multi-Scale Memory-Based Video Deblurring
Bo Ji, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1919-1928
Video deblurring has achieved remarkable progress thanks to the success of deep neural networks. Most methods solve for the deblurring end-to-end with limited i nformation propagation from the video sequence. However, different frame regions exhibit different characteristics and should be provided with corresponding rel evant information. To achieve fine-grained deblurring, we designed a memory bran ch to memorize the blurry-sharp feature pairs in the memory bank, thus providing useful information for the blurry query input. To enrich the memory of our memo

ry bank, we further designed a bidirectional recurrency and multi-scale strategy based on the memory bank. Experimental results demonstrate that our model outperforms other state-of-the-art methods while keeping the model complexity and inference time low. The code is available at https://github.com/jibo27/MemDeblur.

****

SkinningNet: Two-Stream Graph Convolutional Neural Network for Skinning Prediction of Synthetic Characters

Albert Mosella-Montoro, Javier Ruiz-Hidalgo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18593-18602

This work presents SkinningNet, an end-to-end Two-Stream Graph Neural Network architecture that computes skinning weights from an input mesh and its associated skeleton, without making any assumptions on shape class and structure of the provided mesh. Whereas previous methods pre-compute handcrafted features that relate the mesh and the skeleton or assume a fixed topology of the skeleton, the proposed method extracts this information in an end-to-end learnable fashion by jointly learning the best relationship between mesh vertices and skeleton joints. The proposed method exploits the benefits of the novel Multi-Aggregator Graph Convolution that combines the results of different aggregators during the summarizing step of the Message-Passing scheme, helping the operation to generalize for unseen topologies. Experimental results demonstrate the effectiveness of the contributions of our novel architecture, with SkinningNet outperforming current state-of-the-art alternatives.

****

A Scalable Combinatorial Solver for Elastic Geometrically Consistent 3D Shape Matching

Paul Roetzer, Paul Swoboda, Daniel Cremers, Florian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 428-438

We present a scalable combinatorial algorithm for globally optimizing over the space of geometrically consistent mappings between 3D shapes. We use the mathematically elegant formalism proposed by Windheuser et al. (ICCV, 2011) where 3D shape matching was formulated as an integer linear program over the space of orientation-preserving diffeomorphisms. Until now, the resulting formulation had limited practical applicability due to its complicated constraint structure and its large size. We propose a novel primal heuristic coupled with a Lagrange dual problem that is several orders of magnitudes faster compared to previous solvers. This allows us to handle shapes with substantially more triangles than previously solvable. We demonstrate compelling results on diverse datasets, and, even showcase that we can address the challenging setting of matching two partial shapes without availability of complete shapes. Our code is publicly available at http://github.com/paul0noah/sm-comb.

****

Learning Trajectory-Aware Transformer for Video Super-Resolution

Chengxu Liu, Huan Yang, Jianlong Fu, Xueming Qian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5687-5696

Video super-resolution (VSR) aims to restore a sequence of high-resolution (HR) frames from their low-resolution (LR) counterparts. Although some progress has been made, there are grand challenges to effectively utilize temporal dependency in entire video sequences. Existing approaches usually align and aggregate video frames from limited adjacent frames (e.g., 5 or 7 frames), which prevents these approaches from satisfactory results. In this paper, we take one step further to enable effective spatio-temporal learning in videos. We propose a novel Trajectory-aware Transformer for Video Super-Resolution (TTVSR). In particular, we formulate video frames into several pre-aligned trajectories which consist of continuous visual tokens. For a query token, self-attention is only learned on relevant visual tokens along spatio-temporal trajectories. Compared with vanilla vision Transformers, such a design significantly reduces the computational cost and enables Transformers to model long-range features. We further propose a cross-scale feature tokenization module to overcome scale-changing problems that often occur in long-range videos. Experimental results demonstrate the superiority of th

e proposed TTVSR over state-of-the-art models, by extensive quantitative and qua
litative evaluations in four widely-used video super-resolution benchmarks.
****

Differentiable Dynamics for Articulated 3D Human Motion Reconstruction
Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, Cristian Sminchisescu; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2022, pp. 13190-13200
We introduce DiffPhy, a differentiable physics-based model for articulated 3d hu
man motion reconstruction from video. Applications of physics-based reasoning in
 human motion analysis have so far been limited, both by the complexity of const
ructing adequate physical models of articulated human motion, and by the formida
ble challenges of performing stable and efficient inference with physics in the
loop. We jointly address such modeling and inference challenges by proposing an
approach that combines a physically plausible body representation with anatomica
l joint limits, a differentiable physics simulator, and optimization techniques
that ensure good performance and robustness to suboptimal local optima. In contr
ast to several recent methods, our approach readily supports full-body contact i
ncluding interactions with objects in the scene. Most importantly, our model con
nects end-to-end with images, thus supporting direct gradient-based physics opti
mization by means of image-based loss functions. We validate the model by demons
trating that it can accurately reconstruct physically plausible 3d human motion
from monocular video, both on public benchmarks with available 3d ground-truth,
and on videos from the internet.
****

Geometric Structure Preserving Warp for Natural Image Stitching
Peng Du, Jifeng Ning, Jiguang Cui, Shaoli Huang, Xinchao Wang, Jiaxin Wang; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 3688-3696
Preserving geometric structures in the scene plays a vital role in image stitchi
ng. However, most of the existing methods ignore the large-scale layouts reflect
ed by straight lines or curves, decreasing overall stitching quality. To address
 this issue, this work presents a structure-preserving stitching approach that p
roduces images with natural visual effects and less distortion. Our method first
 employs deep learning-based edge detection to extract various types of large-sc
ale edges. Then, the extracted edges are sampled to construct multiple groups of
 triangles to represent geometric structures. Meanwhile, a GEometric Structure p
reserving (GES) energy term is introduced to make these triangles undergo simila
rity transformation. Further, an optimized GES energy term is presented to reaso
nably determine the weights of the sampling points on the geometric structure, a
nd the term is added into the Global Similarity Prior (GSP) stitching model call
ed GES-GSP to achieve a smooth transition between local alignment and geometric
structure preservation. The effectiveness of GES-GSP is validated through compre
hensive experiments on a stitching dataset. The experimental results show that t
he proposed method outperforms several state-of-the-art methods in geometric str
ucture preservation and obtains more natural stitching results. The code and dat
aset are available at https://github.com/flowerDuo/GES-GSP-Stitching.
****

GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping
Omid Taheri, Vasileios Choutas, Michael J. Black, Dimitrios Tzionas; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
022, pp. 13263-13273
Generating digital humans that move realistically has many applications and is w
idely studied, but existing methods focus on the major limbs of the body, ignori
ng the hands and head. Hands have been separately studied, but the focus has bee
n on generating realistic static grasps of objects. To synthesize virtual charac
ters that interact with the world, we need to generate full-body motions and rea
listic hand grasps simultaneously. Both sub-problems are challenging on their ow
n and, together, the state space of poses is significantly larger, the scales of
 hand and body motions differ, and the whole-body posture and the hand grasp mus
t agree, satisfy physical constraints, and be plausible. Additionally, the head

is involved because the avatar must look at the object to interact with it. For the first time, we address the problem of generating full-body, hand and head motions of an avatar grasping an unknown object. As input, our method, called GOAL, takes a 3D object, its pose, and a starting 3D body pose and shape. GOAL outputs a sequence of whole-body poses using two novel networks. First, GNet generates a goal whole-body grasp with a realistic body, head, arm, and hand pose, as well as hand-object contact. Second, MNet generates the motion between the starting and goal pose. This is challenging, as it requires the avatar to walk towards the object with foot-ground contact, orient the head towards it, reach out, and grasp it with a realistic hand pose and hand-object contact. To achieve this the networks exploit a representation that combines SMPL-X body parameters and 3D vertex offsets. We train and evaluate GOAL, both qualitatively and quantitatively, on the GRAB dataset. Results show that GOAL generalizes well to unseen objects, outperforming baselines. A perceptual study shows that GOAL's generated motions approach the realism of GRAB's ground truth. GOAL takes a step towards generating realistic full-body object grasping motion. Our models and code are available at https://goal.is.tue.mpg.de.

****

Multi-Robot Active Mapping via Neural Bipartite Graph Matching

Kai Ye, Siyan Dong, Qingnan Fan, He Wang, Li Yi, Fei Xia, Jue Wang, Baoquan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14839-14848

We study the problem of multi-robot active mapping, which aims for complete scene map construction in minimum time steps. The key to this problem lies in the goal position estimation to enable more efficient robot movements. Previous approaches either choose the frontier as the goal position via a myopic solution that hinders the time efficiency, or maximize the long-term value via reinforcement learning to directly regress the goal position, but does not guarantee the complete map construction. In this paper, we propose a novel algorithm, namely NeuralCoMapping, which takes advantage of both approaches. We reduce the problem to bipartite graph matching, which establishes the node correspondences between two graphs, denoting robots and frontiers. We introduce a multiplex graph neural network (mGNN) that learns the neural distance to fill the affinity matrix for more effective graph matching. We optimize the mGNN with a differentiable linear assignment layer by maximizing the long-term values that favor time efficiency and map completeness via reinforcement learning. We compare our algorithm with several state-of-the-art multi-robot active mapping approaches and adapted reinforcement-learning baselines. Experimental results demonstrate the superior performance and exceptional generalization ability of our algorithm on various indoor scenes and unseen number of robots, when only trained with 9 indoor scenes.

****

Adversarial Texture for Fooling Person Detectors in the Physical World

Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, Xiaolin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13307-13316

Nowadays, cameras equipped with AI systems can capture and analyze images to detect people automatically. However, the AI system can make mistakes when receiving deliberately designed patterns in the real world, i.e., physical adversarial examples. Prior works have shown that it is possible to print adversarial patches on clothes to evade DNN-based person detectors. However, these adversarial examples could have catastrophic drops in the attack success rate when the viewing angle (i.e., the camera's angle towards the object) changes. To perform a multi-angle attack, we propose Adversarial Texture (AdvTexture). AdvTexture can cover clothes with arbitrary shapes so that people wearing such clothes can hide from person detectors from different viewing angles. We propose a generative method, named Toroidal-Cropping-based Expandable Generative Attack (TC-EGA), to craft AdvTexture with repetitive structures. We printed several pieces of cloth with AdvTexure and then made T-shirts, skirts, and dresses in the physical world. Experiments showed that these clothes could fool person detectors in the physical world.

****
Focal Length and Object Pose Estimation via Render and Compare
Georgy Ponimatkin, Yann Labbé, Bryan Russell, Mathieu Aubry, Josef Sivic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3825-3834
We introduce FocalPose, a neural render-and-compare method for jointly estimating the camera-object 6D pose and camera focal length given a single RGB input image depicting a known object. The contributions of this work are twofold. First, we derive a focal length update rule that extends an existing state-of-the-art render-and-compare 6D pose estimator to address the joint estimation task. Second, we investigate several different loss functions for jointly estimating the object pose and focal length. We find that a combination of direct focal length regression with a reprojection loss disentangling the contribution of translation, rotation, and focal length leads to improved results. We show results on three challenging benchmark datasets that depict known 3D models in uncontrolled settings. We demonstrate that our focal length and 6D pose estimates have lower error than the existing state-of-the-art methods.
****
TO-FLOW: Efficient Continuous Normalizing Flows With Temporal Optimization Adjoint With Moving Speed
Shian Du, Yihong Luo, Wei Chen, Jian Xu, Delu Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12570-12580
Continuous normalizing flows (CNFs) construct invertible mappings between an arbitrary complex distribution and an isotropic Gaussian distribution using Neural Ordinary Differential Equations (neural ODEs). It has not been tractable on large datasets due to the incremental complexity of the neural ODE training. Optimal Transport theory has been applied to regularize the dynamics of the ODE to speed up training in recent works. In this paper, a temporal optimization is proposed by optimizing the evolutionary time for forward propagation of the neural ODE training. In this appoach, we optimize the network weights of the CNF alternately with evolutionary time by coordinate descent. Further with temporal regularization, stability of the evolution is ensured. This approach can be used in conjunction with the original regularization approach. We have experimentally demonstrated that the proposed approach can significantly accelerate training without sacrifying performance over baseline models.
****
Arbitrary-Scale Image Synthesis
Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Radu Timofte, Martin Danelljan, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11533-11542
Positional encodings have enabled recent works to train a single adversarial network that can generate images of different scales. However, these approaches are either limited to a set of discrete scales or struggle to maintain good perceptual quality at the scales for which the model is not trained explicitly. We propose the design of scale-consistent positional encodings invariant to our generator's layers transformations. This enables the generation of arbitrary-scale images even at scales unseen during training. Moreover, we incorporate novel inter-scale augmentations into our pipeline and partial generation training to facilitate the synthesis of consistent images at arbitrary scales. Lastly, we show competitive results for a continuum of scales on various commonly used datasets for image synthesis.
****
Cross-Modal Representation Learning for Zero-Shot Action Recognition
Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, Linjie Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19978-19988
We present a cross-modal Transformer-based framework, which jointly encodes video data and text labels for zero-shot action recognition (ZSAR). Our model employs a conceptually new pipeline by which visual representations are learned in con

junction with visual-semantic associations in an end-to-end manner. The model de
sign provides a natural mechanism for visual and semantic representations to be
learned in a shared knowledge space, whereby it encourages the learned visual em
bedding to be discriminative and more semantically consistent. In zero-shot infe
rence, we devise a simple semantic transfer scheme that embeds semantic relatedn
ess information between seen and unseen classes to composite unseen visual proto
types. Accordingly, the discriminative features in the visual structure could be
 preserved and exploited to alleviate the typical zero-shot issues of informatio
n loss, semantic gap, and the hubness problem. Under a rigorous zero-shot settin
g of not pre-training on additional datasets, the experiment results show our mo
del considerably improves upon the state of the arts in ZSAR, reaching encouragi
ng top-1 accuracy on UCF101, HMDB51, and ActivityNet benchmark datasets. Code wi
ll be made available.
****

Conditional Prompt Learning for Vision-Language Models
Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 168
16-16825
With the rise of powerful pre-trained vision-language models like CLIP, it becom
es essential to investigate ways to adapt these models to downstream datasets. A
 recently proposed method named Context Optimization (CoOp) introduces the conce
pt of prompt learning---a recent trend in NLP---to the vision domain for adaptin
g pre-trained vision-language models. Specifically, CoOp turns context words in
a prompt into a set of learnable vectors and, with only a few labeled images for
 learning, can achieve huge improvements over intensively-tuned manual prompts.
In our study we identify a critical problem of CoOp: the learned context is not
generalizable to wider unseen classes within the same dataset, suggesting that C
oOp overfits base classes observed during training. To address the problem, we p
ropose Conditional Context Optimization (CoCoOp), which extends CoOp by further
learning a lightweight neural network to generate for each image an input-condit
ional token (vector). Compared to CoOp's static prompts, our dynamic prompts ada
pt to each instance and are thus less sensitive to class shift. Extensive experi
ments show that CoCoOp generalizes much better than CoOp to unseen classes, even
 showing promising transferability beyond a single dataset; and yields stronger
domain generalization performance as well. Code is available at https://github.c
om/KaiyangZhou/CoOp.
****

Graph Sampling Based Deep Metric Learning for Generalizable Person Re-Identifica
tion
Shengcai Liao, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2022, pp. 7359-7368
Recent studies show that, both explicit deep feature matching as well as large-s
cale and diverse training data can significantly improve the generalization of p
erson re-identification. However, the efficiency of learning deep matchers on la
rge-scale data has not yet been adequately studied. Though learning with classif
ication parameters or class memory is a popular way, it incurs large memory and
computational costs. In contrast, pairwise deep metric learning within mini batc
hes would be a better choice. However, the most popular random sampling method,
the well-known PK sampler, is not informative and efficient for deep metric lear
ning. Though online hard example mining has improved the learning efficiency to
some extent, the mining in mini batches after random sampling is still limited.
This inspires us to explore the use of hard example mining earlier, in the data
sampling stage. To do so, in this paper, we propose an efficient mini-batch samp
ling method, called graph sampling (GS), for large-scale deep metric learning. T
he basic idea is to build a nearest neighbor relationship graph for all classes
at the beginning of each epoch. Then, each mini batch is composed of a randomly
selected class and its nearest neighboring classes so as to provide informative
and challenging examples for learning. Together with an adapted competitive base
line, we improve the state of the art in generalizable person re-identification
significantly, by 25.1% in Rank-1 on MSMT17 when trained on RandPerson. Besides,

the proposed method also outperforms the competitive baseline, by 6.8% in Rank-1 on CUHK03-NP when trained on MSMT17. Meanwhile, the training time is significantly reduced, from 25.4 hours to 2 hours when trained on RandPerson with 8,000 identities. Code is available at https://github.com/ShengcaiLiao/QAConv.
****

Retrieval-Based Spatially Adaptive Normalization for Semantic Image Synthesis
Yupeng Shi, Xiao Liu, Yuxiang Wei, Zhongqin Wu, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11224-11233
Semantic image synthesis is a challenging task with many practical applications. Albeit remarkable progress has been made in semantic image synthesis with spatially-adaptive normalization and existing methods normalize the feature activations under the coarse-level guidance (e.g., semantic class). However, different parts of a semantic object (e.g., wheel and window of car) are quite different in structures and textures, making blurry synthesis results usually inevitable due to the missing of fine-grained guidance. In this paper, we propose a novel normalization module, termed as REtrieval-based Spatially AdaptIve normaLization (RESAIL), for introducing pixel level fine-grained guidance to the normalization architecture. Specifically, we first present a retrieval paradigm by finding a content patch of the same semantic class from training set with the most similar shape to each test semantic mask. Then, RESAIL is presented to use the retrieved patch for guiding the feature normalization of corresponding region, and can provide pixel level fine-grained guidance, thereby greatly mitigating blurry synthesis results. Moreover, distorted ground-truth images are also utilized as alternatives of retrieval-based guidance for feature normalization, further benefiting model training and improving visual quality of generated images. Experiments on several challenging datasets show that our RESAIL performs favorably against state-of-the-arts in terms of quantitative metrics, visual quality, and subjective evaluation. The source code and pre-trained models will be publicly available.
****

Undoing the Damage of Label Shift for Cross-Domain Semantic Segmentation
Yahao Liu, Jinhong Deng, Jiale Tao, Tong Chu, Lixin Duan, Wen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7042-7052
Existing works typically treat cross-domain semantic segmentation(CDSS) as a data distribution mismatch problem and focus on aligning the marginal distribution or conditional distribution. However, the label shift issue is unfortunately overlooked, which actually commonly exists in the CDSS task, and often causes a classifier bias in the learnt model. In this paper, we give an in-depth analysis and show that the damage of label shift can be overcome by aligning the data conditional distribution and correcting the posterior probability. To this end, we propose a novel approach to undo the damage of the label shift problem in CDSS. In implementation, we adopt class-level feature alignment for conditional distribution alignment, as well as two simple yet effective methods to rectify the classifier bias from source to target by remolding the classifier predictions. We conduct extensive experiments on the benchmark datasets of urban scenes, including GTA5 to Cityscapes and SYNTHIA to Cityscapes, where our proposed approach outperforms previous methods by a large margin. For instance, our model equipped with a self-training strategy reaches 59.3% mIoU on GTA5 to Cityscapes, pushing to a new state-of-the-art. The code will be available at https://github.com/manmanjun/Undoing_UDA.
****

GPV-Pose: Category-Level Object Pose Estimation via Geometry-Guided Point-Wise Voting
Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6781-6791
While 6D object pose estimation has recently made a huge leap forward, most methods can still only handle a single or a handful of different objects, which limits their applications. To circumvent this problem, category-level object pose es

timation has recently been revamped, which aims at predicting the 6D pose as wel l as the 3D metric size for previously unseen instances from a given set of obje ct classes. This is, however, a much more challenging task due to severe intra-c lass shape variations. To address this issue, we propose GPV-Pose, a novel frame work for robust category-level pose estimation, harnessing geometric insights to enhance the learning of category-level pose-sensitive features. First, we intro duce a decoupled confidence-driven rotation representation, which allows geometr y-aware recovery of the associated rotation matrix. Second, we propose a novel g eometry-guided point-wise voting paradigm for robust retrieval of the 3D object bounding box. Finally, leveraging these different output streams, we can enforce several geometric consistency terms, further increasing performance, especially for non-symmetric categories. GPV-Pose produces superior results to state-of-th e-art competitors on common public benchmarks, whilst almost achieving real-time inference speed at 20 FPS. Code and trained models will be made publicly availa ble.

****

Dynamic 3D Gaze From Afar: Deep Gaze Estimation From Temporal Eye-Head-Body Coor dination

Soma Nonaka, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2192-2201

We introduce a novel method and dataset for 3D gaze estimation of a freely movin g person from a distance, typically in surveillance views. Eyes cannot be clearl y seen in such cases due to occlusion and lacking resolution. Existing gaze esti mation methods suffer or fall back to approximating gaze with head pose as they primarily rely on clear, close-up views of the eyes. Our key idea is to instead leverage the intrinsic gaze, head, and body coordination of people. Our method f ormulates gaze estimation as Bayesian prediction given temporal estimates of hea d and body orientations which can be reliably estimated from a far. We model the head and body orientation likelihoods and the conditional prior of gaze directi on on those with separate neural networks which are then cascaded to output the 3D gaze direction. We introduce an extensive new dataset that consists of survei llance videos annotated with 3D gaze directions captured in 5 indoor and outdoor scenes. Experimental results on this and other datasets validate the accuracy o f our method and demonstrate that gaze can be accurately estimated from a typica l surveillance distance even when the person's face is not visible to the camera .

****

Expressive Talking Head Generation With Granular Audio-Visual Control

Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, Jingdong Wang; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3387-3396

Generating expressive talking heads is essential for creating virtual humans. Ho wever, existing one- or few-shot methods focus on lip-sync and head motion, igno ring the emotional expressions that make talking faces realistic. In this paper, we propose the Granularly Controlled Audio-Visual Talking Heads (GC-AVT), which controls lip movements, head poses, and facial expressions of a talking head in a granular manner. Our insight is to decouple the audio-visual driving sources through prior-based pre-processing designs. Detailedly, we disassemble the drivi ng image into three complementary parts including: 1) a cropped mouth that facil itates lip-sync; 2) a masked head that implicitly learns pose; and 3) the upper face which works corporately and complementarily with a time-shifted mouth to co ntribute the expression. Interestingly, the encoded features from the three sour ces are integrally balanced through reconstruction training. Extensive experimen ts show that our method generates expressive faces with not only synced mouth sh apes, controllable poses, but precisely animated emotional expressions as well.

****

Trustworthy Long-Tailed Classification

Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, Changqing Zhang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 6970-6979

Classification on long-tailed distributed data is a challenging problem, which s
uffers from serious class-imbalance and accordingly unpromising performance espe
cially on tail classes. Recently, the ensembling based methods achieve the state
-of-the-art performance and show great potential. However, there are two limitat
ions for current methods. First, their predictions are not trustworthy for failu
re-sensitive applications. This is especially harmful for the tail classes where
 the wrong predictions is basically frequent. Second, they assign unified number
s of experts to all samples, which is redundant for easy samples with excessive
computational cost. To address these issues, we propose a Trustworthy Long-taile
d Classification (TLC) method to jointly conduct classification and uncertainty
estimation to identify hard samples in a multi-expert framework. Our TLC obtains
 the evidence-based uncertainty (EvU) and evidence for each expert, and then com
bines these uncertainties and evidences under the Dempster-Shafer Evidence Theor
y (DST). Moreover, we propose a dynamic expert engagement to reduce the number o
f engaged experts for easy samples and achieve efficiency while maintaining prom
ising performances. Finally, we conduct comprehensive experiments on the tasks o
f classification, tail detection, OOD detection and failure prediction. The expe
rimental results show that the proposed TLC outperforms existing methods and is
trustworthy with reliable uncertainty.
****

Primitive3D: 3D Object Dataset Synthesis From Randomly Assembled Primitives
Xinke Li, Henghui Ding, Zekun Tong, Yuwei Wu, Yeow Meng Chee; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 15947-15957
Numerous advancements of deep learning can be attributed to access to large-scal
e and well-annotated datasets. However, such a dataset is prohibitively expensiv
e in 3D computer vision due to the substantial collection cost. To alleviate thi
s issue, we propose a cost-effective method for automatically generating a large
 amount of 3D objects with annotations. In particular, we synthesize objects sim
ply by assembling multiple random primitives. These objects are thus auto-annota
ted with part-based labels originating from primitives. This allows us to perfor
m multi-task learning by combining the supervised segmentation with unsupervised
 reconstruction. Considering the large overhead of learning on the generated dat
aset, we further propose a dataset distillation strategy to remove redundant sam
ples regarding a target dataset. We conduct extensive experiments for the downst
ream tasks of 3D object classification. The results indicate that our dataset, t
ogether with multi-task pretraining on its annotations, achieves the best perfor
mance compared to other commonly used datasets. Further study suggests that our
strategy can improve the model performance by pretraining and fine-tuning scheme
, especially for a dataset with a small scale. In addition, pretraining with the
 proposed dataset distillation method can save 86% of the pretraining time with
negligible performance degradation. We expect that our attempt provides a new da
ta-centric perspective for training 3D deep models.
****

Mix and Localize: Localizing Sound Sources in Mixtures
Xixi Hu, Ziyang Chen, Andrew Owens; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 10483-10492
We present a method for simultaneously localizing multiple sound sources within
a visual scene. This task requires a model to both group a sound mixture into in
dividual sources, and to associate them with a visual signal. Our method jointly
 solves both tasks at once, using a formulation inspired by the contrastive rand
om walk of Jabri et al. We create a graph in which images and separated sounds e
ach correspond to nodes, and train a random walker to transition between nodes f
rom different modalities with high return probability. The transition probabilit
ies for this walk are determined by an audio-visual similarity metric that is le
arned by our model. We show through experiments with musical instruments and hum
an speech that our model can successfully localize multiple sounds, outperformin
g other self-supervised methods.
****

FisherMatch: Semi-Supervised Rotation Regression via Entropy-Based Filtering

Yingda Yin, Yingcheng Cai, He Wang, Baoquan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11164-11173

Estimating the 3DoF rotation from a single RGB image is an important yet challenging problem. Recent works achieve good performance relying on a large amount of expensive-to-obtain labeled data. To reduce the amount of supervision, we for the first time propose a general framework, FisherMatch, for semi-supervised rotation regression, without assuming any domain-specific knowledge or paired data. Inspired by the popular semi-supervised approach, FixMatch, we propose to leverage pseudo label filtering to facilitate the information flow from labeled data to unlabeled data in a teacher-student mutual learning framework. However, incorporating the pseudo label filtering mechanism into semi-supervised rotation regression is highly non-trivial, mainly due to the lack of a reliable confidence measure for rotation prediction. In this work, we propose to leverage matrix Fisher distribution to build a probabilistic model of rotation and devise a matrix Fisher-based regressor for jointly predicting rotation along with its prediction uncertainty. We then propose to use the entropy of the predicted distribution as a confidence measure, which enables us to perform pseudo label filtering for rotation regression. For supervising such distribution-like pseudo labels, we further investigate the problem of how to enforce loss between two matrix Fisher distributions. Our extensive experiments show that our method can work well even under very low labeled data ratios on different benchmarks, achieving significant and consistent performance improvement over supervised learning and other semi-supervised learning baselines. Our project page is at https://yd-yin.github.io/FisherMatch.

****

NPBG++: Accelerating Neural Point-Based Graphics

Ruslan Rakhimov, Andrei-Timotei Ardelean, Victor Lempitsky, Evgeny Burnaev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15969-15979

We present a new system (NPBG++) for the novel view synthesis (NVS) task that achieves high rendering realism with low scene fitting time. Our method efficiently leverages the multiview observations and the point cloud of a static scene to predict a neural descriptor for each point, improving upon the pipeline of Neural Point-Based Graphics in several important ways. By predicting the descriptors with a single pass through the source images, we lift the requirement of per-scene optimization while also making the neural descriptors view-dependent and more suitable for scenes with strong non-Lambertian effects. In our comparisons, the proposed system outperforms previous NVS approaches in terms of fitting and rendering runtimes while producing images of similar quality.

****

SphericGAN: Semi-Supervised Hyper-Spherical Generative Adversarial Networks for Fine-Grained Image Synthesis

Tianyi Chen, Yunfei Zhang, Xiaoyang Huo, Si Wu, Yong Xu, Hau San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10001-10010

Generative Adversarial Network (GAN)-based models have greatly facilitated image synthesis. However, the model performance may be degraded when applied to fine-grained data, due to limited training samples and subtle distinction among categories. Different from generic GANs, we address the issue from a new perspective of discovering and utilizing the underlying structure of real data to explicitly regularize the spatial organization of latent space. To reduce the dependence of generative models on labeled data, we propose a semi-supervised hyper-spherical GAN for class-conditional fine-grained image generation, and our model is referred to as SphericGAN. By projecting random vectors drawn from a prior distribution onto a hyper-sphere, we can model more complex distributions, while at the same time the similarity between the resulting latent vectors depends only on the angle, but not on their magnitudes. On the other hand, we also incorporate a mapping network to map real images onto the hyper-sphere, and match latent vectors with the underlying structure of real data via real-fake cluster alignment. As

a result, we obtain a spatially organized latent space, which is useful for capturing class-independent variation factors. The experimental results suggest that our SphericGAN achieves state-of-the-art performance in synthesizing high-fidelity images with precise class semantics.
****

HairMapper: Removing Hair From Portraits Using GANs
Yiqian Wu, Yong-Liang Yang, Xiaogang Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4227-4236
Removing hair from portrait images is challenging due to the complex occlusions between hair and face, as well as the lack of paired portrait data with/without hair. To this end, we present a dataset and a baseline method for removing hair from portrait images using generative adversarial networks (GANs). Our core idea is to train a fully connected network HairMapper to find the direction of hair removal in the latent space of StyleGAN for the training stage. We develop a new separation boundary and diffuse method to generate paired training data for males, and a novel "female-male-bald" pipeline for paired data of females. Experiments show that our method can naturally deal with portrait images with variations on gender, age, etc. We validate the superior performance of our method by comparing it to state-of-the-art methods through extensive experiments and user studies. We also demonstrate its applications in hair design and 3D face reconstruction.
****

Affine Medical Image Registration With Coarse-To-Fine Vision Transformer
Tony C. W. Mok, Albert C. S. Chung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20835-20844
Affine registration is indispensable in a comprehensive medical image registration pipeline. However, only a few studies focus on fast and robust affine registration algorithms. Most of these studies utilize convolutional neural networks (CNNs) to learn joint affine and non-parametric registration, while the standalone performance of the affine subnetwork is less explored. Moreover, existing CNN-based affine registration approaches focus either on the local misalignment or the global orientation and position of the input to predict the affine transformation matrix, which are sensitive to spatial initialization and exhibit limited generalizability apart from the training dataset. In this paper, we present a fast and robust learning-based algorithm, Coarse-to-Fine Vision Transformer (C2FViT), for 3D affine medical image registration. Our method naturally leverages the global connectivity and locality of the convolutional vision transformer and the multi-resolution strategy to learn the global affine registration. We evaluate our method on 3D brain atlas registration and template-matching normalization. Comprehensive results demonstrate that our method is superior to the existing CNNs-based affine registration methods in terms of registration accuracy, robustness and generalizability while preserving the runtime advantage of the learning-based methods. The source code is available at https://github.com/cwmok/C2FViT.
****

SMPL-A: Modeling Person-Specific Deformable Anatomy
Hengtao Guo, Benjamin Planche, Meng Zheng, Srikrishna Karanam, Terrence Chen, Ziyan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20814-20823
A variety of diagnostic and therapeutic protocols rely on locating in vivo target anatomical structures, which can be obtained from medical scans. However, organs move and deform as the patient changes his/her pose. In order to obtain accurate target location information, clinicians have to either conduct frequent intraoperative scans, resulting in higher exposition of patients to radiations, or adopt proxy procedures (e.g., creating and using custom molds to keep patients in the exact same pose during both preoperative organ scanning and subsequent treatment. Such custom proxy methods are typically sub-optimal, constraining the clinicians and costing precious time and money to the patients. To the best of our knowledge, this work is the first to present a learning-based approach to estimate the patient's internal organ deformation for arbitrary human poses in order to assist with radiotherapy and similar medical protocols. The underlying method

first leverages medical scans to learn a patient-specific representation that po tentially encodes the organ's shape and elastic properties. During inference, gi ven the patient's current body pose information and the organ's representation e xtracted from previous medical scans, our method can estimate their current orga n deformation to offer guidance to clinicians. We conduct experiments on a well- sized dataset which is augmented through real clinical data using finite element modeling. Our results suggest that pose-dependent organ deformation can be lear ned through a point cloud autoencoder conditioned on the parametric pose input. We hope that this work can be a starting point for future research towards closi ng the loop between human mesh recovery and anatomical reconstruction, with appl ications beyond the medical domain.

****

Image Dehazing Transformer With Transmission-Aware 3D Position Embedding

Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, Chongyi Li; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5812-5820

Despite single image dehazing has been made promising progress with Convolutiona l Neural Networks (CNNs), the inherent equivariance and locality of convolution still bottleneck dehazing performance. Though Transformer has occupied various c omputer vision tasks, directly leveraging Transformer for image dehazing is chal lenging: 1) it tends to result in ambiguous and coarse details that are undesire d for image reconstruction; 2) previous position embedding of Transformer is pro vided in logic or spatial position order that neglects the variational haze dens ities, which results in the sub-optimal dehazing performance. The key insight of this study is to investigate how to combine CNN and Transformer for image dehaz ing. To solve the feature inconsistency issue between Transformer and CNN, we pr opose to modulate CNN features via learning modulation matrices (i.e., coefficie nt matrix and bias matrix) conditioned on Transformer features instead of simple feature addition or concatenation. The feature modulation naturally inherits th e global context modeling capability of Transformer and the local representation capability of CNN. We bring a haze density-related prior into Transformer via a novel transmission-aware 3D position embedding module, which not only provides the relative position but also suggests the haze density of different spatial re gions. Extensive experiments demonstrate that our method, DeHamer, attains state -of-the-art performance on several image dehazing benchmarks.

****

Out-of-Distribution Generalization With Causal Invariant Transformations

Ruoyu Wang, Mingyang Yi, Zhitang Chen, Shengyu Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 375-385

In real-world applications, it is important and desirable to learn a model that performs well on out-of-distribution (OOD) data. Recently, causality has become a powerful tool to tackle the OOD generalization problem, with the idea resting on the causal mechanism that is invariant across domains of interest. To leverag e the generally unknown causal mechanism, existing works assume a linear form of causal feature or require sufficiently many and diverse training domains, which are usually restrictive in practice. In this work, we obviate these assumptions and tackle the OOD problem without explicitly recovering the causal feature. Ou r approach is based on transformations that modify the non-causal feature but le ave the causal part unchanged, which can be either obtained from prior knowledge or learned from the training data in the multi-domain scenario. Under the setti ng of invariant causal mechanism, we theoretically show that if all such transfo rmations are available, then we can learn a minimax optimal model across the dom ains using only single domain data. Noticing that knowing a complete set of thes e causal invariant transformations may be impractical, we further show that it s uffices to know only a subset of these transformations. Based on the theoretical findings, a regularized training procedure is proposed to improve the OOD gener alization capability. Extensive experimental results on both synthetic and real datasets verify the effectiveness of the proposed algorithm, even with only a fe w causal invariant transformations.

****

Panoptic-PHNet: Towards Real-Time and High-Precision LiDAR Panoptic Segmentation via Clustering Pseudo Heatmap
Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, Dan Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11809-11818
As a rising task, panoptic segmentation is faced with challenges in both semantic segmentation and instance segmentation. However, in terms of speed and accuracy, existing LiDAR methods in the field are still limited. In this paper, we propose a fast and high-performance LiDAR-based framework, referred to as Panoptic-PHNet, with three attractive aspects: 1) We introduce a clustering pseudo heatmap as a new paradigm, which, followed by a center grouping module, yields instance centers for efficient clustering without object-level learning tasks. 2) A knn-transformer module is proposed to model the interaction among foreground points for accurate offset regression. 3) For backbone design, we fuse the fine-grained voxel features and the 2D Bird's Eye View (BEV) features with different receptive fields to utilize both detailed and global information. Extensive experiments on both SemanticKITTI dataset and nuScenes dataset show that our Panoptic-PHNet surpasses state-of-the-art methods by remarkable margins with a real-time speed. We achieve the 1st place on the public leaderboard of SemanticKITTI and leading performance on the recently released leaderboard of nuScenes.
****
Dual-Key Multimodal Backdoors for Visual Question Answering
Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, Susmit Jha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15375-15385
The success of deep learning has enabled advances in multimodal tasks that require non-trivial fusion of multiple input domains. Although multimodal models have shown potential in many problems, their increased complexity makes them more vulnerable to attacks. A Backdoor (or Trojan) attack is a class of security vulnerability wherein an attacker embeds a malicious secret behavior into a network (e.g. targeted misclassification) that is activated when an attacker-specified trigger is added to an input. In this work, we show that multimodal networks are vulnerable to a novel type of attack that we refer to as Dual-Key Multimodal Backdoors. This attack exploits the complex fusion mechanisms used by state-of-the-art networks to embed backdoors that are both effective and stealthy. Instead of using a single trigger, the proposed attack embeds a trigger in each of the input modalities and activates the malicious behavior only when both the triggers are present. We present an extensive study of multimodal backdoors on the Visual Question Answering (VQA) task with multiple architectures and visual feature backbones. A major challenge in embedding backdoors in VQA models is that most models use visual features extracted from a fixed pretrained object detector. This is challenging for the attacker as the detector can distort or ignore the visual trigger entirely, which leads to models where backdoors are over-reliant on the language trigger. We tackle this problem by proposing a visual trigger optimization strategy designed for pretrained object detectors. Through this method, we create Dual-Key Backdoors with over a 98% attack success rate while only poisoning 1% of the training data. Finally, we release TrojVQA, a large collection of clean and trojan VQA models to enable research in defending against multimodal backdoors.
****
A Differentiable Two-Stage Alignment Scheme for Burst Image Reconstruction With Large Shift
Shi Guo, Xi Yang, Jianqi Ma, Gaofeng Ren, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17472-17481
Denoising and demosaicking are two essential steps to reconstruct a clean full-color image from the raw data. Recently, joint denoising and demosaicking (JDD) for burst images, namely JDD-B, has attracted much attention by using multiple raw images captured in a short time to reconstruct a single high-quality image. One key challenge of JDD-B lies in the robust alignment of image frames. State-of-

the-art alignment methods in feature domain cannot effectively utilize the tempo ral information of burst images, where large shifts commonly exist due to camera and object motion. In addition, the higher resolution (e.g., 4K) of modern imag ing devices results in larger displacement between frames. To address these chal lenges, we design a differentiable two-stage alignment scheme sequentially in pa tch and pixel level for effective JDD-B. The input burst images are firstly alig ned in the patch level by using a differentiable progressive block matching meth od, which can estimate the offset between distant frames with small computationa l cost. Then we perform implicit pixel-wise alignment in full-resolution feature domain to refine the alignment results. The two stages are jointly trained in a n end-to-end manner. Extensive experiments demonstrate the significant improveme nt of our method over existing JDD-B methods. Codes are available at https://git hub.com/GuoShi28/2StageAlign.

****

Unifying Panoptic Segmentation for Autonomous Driving

Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, Csaba Be leznai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 21351-21360

This paper aims to improve panoptic segmentation for real-world applications in three ways. First, we present a label policy that unifies four of the most popul ar panoptic segmentation datasets for autonomous driving. We also clean up label confusion by adding the new vehicle labels pickup and van. Full relabeling info rmation for the popular Mapillary Vistas, IDD, and Cityscapes dataset are provid ed to add these new labels to existing setups. Second, we introduce Wilddash2 (W D2), a new dataset and public benchmark service for panoptic segmentation. The d ataset consists of more than 5000 unique driving scenes from all over the world with a focus on visually challenging scenes, such as diverse weather conditions, lighting situations, and camera characteristics. We showcase experimental visua l hazard classifiers which help to pre-filter challenging frames during dataset creation. Finally, to characterize the robustness of algorithms in out-of-distri bution situations, we introduce hazard-aware and negative testing for panoptic s egmentation as well as statistical significance calculations that increase confi dence for both concepts. Additionally, we present a novel technique for visualiz ing panoptic segmentation errors. Our experiments show the negative impact of vi sual hazards on panoptic segmentation quality. Additional data from the WD2 data set improves performance for visually challenging scenes and thus robustness in real-world scenarios.

****

Learning Motion-Dependent Appearance for High-Fidelity Rendering of Dynamic Huma ns From a Single Camera

Jae Shin Yoon, Duygu Ceylan, Tuanfeng Y. Wang, Jingwan Lu, Jimei Yang, Zhixin Sh u, Hyun Soo Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3407-3417

Appearance of dressed humans undergoes a complex geometric transformation induce d not only by the static pose but also by its dynamics, i.e., there exists a num ber of cloth geometric configurations given a pose depending on the way it has m oved. Such appearance modeling conditioned on motion has been largely neglected in existing human rendering methods, resulting in rendering of physically implau sible motion. A key challenge of learning the dynamics of the appearance lies in the requirement of a prohibitively large amount of observations. In this paper, we present a compact motion representation by enforcing equivariance---a repres entation is expected to be transformed in the way that the pose is transformed. We model an equivariant encoder that can generate the generalizable representati on from the spatial and temporal derivatives of the 3D body surface. This learne d representation is decoded by a compositional multi-task decoder that renders h igh fidelity time-varying appearance. Our experiments show that our method can g enerate a temporally coherent video of dynamic humans for unseen body poses and novel views given a single view video.

****

On the Road to Online Adaptation for Semantic Image Segmentation

Riccardo Volpi, Pau De Jorge, Diane Larlus, Gabriela Csurka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19184-19195

We propose a new problem formulation and a corresponding evaluation framework to advance research on unsupervised domain adaptation for semantic image segmentation. The overall goal is fostering the development of adaptive learning systems that will continuously learn, without supervision, in ever-changing environments. Typical protocols that study adaptation algorithms for segmentation models are limited to few domains, adaptation happens offline, and human intervention is generally required, at least to annotate data for hyper-parameter tuning. We argue that such constraints are incompatible with algorithms that can continuously adapt to different real-world situations. To address this, we propose a protocol where models need to learn online, from sequences of temporally correlated images, requiring continuous, frame-by-frame adaptation. We accompany this new protocol with a variety of baselines to tackle the proposed formulation, as well as an extensive analysis of their behaviors, which can serve as a starting point for future research.

****

Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes

Jon Donnelly, Alina Jade Barnett, Chaofan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10265-10275

We present a deformable prototypical part network (Deformable ProtoPNet), an interpretable image classifier that integrates the power of deep learning and the interpretability of case-based reasoning. This model classifies input images by comparing them with prototypes learned during training, yielding explanations in the form of "this looks like that." However, while previous methods use spatially rigid prototypes, we address this shortcoming by proposing spatially flexible prototypes. Each prototype is made up of several prototypical parts that adaptively change their relative spatial positions depending on the input image. Consequently, a Deformable ProtoPNet can explicitly capture pose variations and context, improving both model accuracy and the richness of explanations provided. Compared to other case-based interpretable models using prototypes, our approach achieves state-of-the-art accuracy and gives an explanation with greater context. The code is available at https://github.com/jdonnelly36/Deformable-ProtoPNet.

****

Context-Aware Sequence Alignment Using 4D Skeletal Augmentation

Taein Kwon, Bugra Tekin, Siyu Tang, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8172-8182

Temporal alignment of fine-grained human actions in videos is important for numerous applications in computer vision, robotics, and mixed reality. State-of-the-art methods directly learn image-based embedding space by leveraging powerful deep convolutional neural networks. While being straightforward, their results are far from satisfactory, the aligned videos exhibit severe temporal discontinuity without additional post-processing steps. The recent advancements in human body and hand pose estimation in the wild promise new ways of addressing the task of human action alignment in videos. In this work, based on off-the-shelf human pose estimators, we propose a novel context-aware self-supervised learning architecture to align sequences of actions. We name it CASA. Specifically, CASA employs self-attention and cross-attention mechanisms to incorporate the spatial and temporal context of human actions, which can solve the temporal discontinuity problem. Moreover, we introduce a self-supervised learning scheme that is empowered by novel 4D augmentation techniques for 3D skeleton representations. We systematically evaluate the key components of our method. Our experiments on three public datasets demonstrate CASA significantly improves phase progress and Kendall's Tau scores over the previous state-of-the-art methods.

****

Perturbed and Strict Mean Teachers for Semi-Supervised Semantic Segmentation

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, Gustavo

Carneiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4258-4267
Consistency learning using input image, feature, or network perturbations has shown remarkable results in semi-supervised semantic segmentation, but this approach can be seriously affected by inaccurate predictions of unlabelled training images. There are two consequences of these inaccurate predictions: 1) the training based on the "strict" cross-entropy (CE) loss can easily overfit prediction mistakes, leading to confirmation bias; and 2) the perturbations applied to these inaccurate predictions will use potentially erroneous predictions as training signals, degrading consistency learning. In this paper, we address the prediction accuracy problem of consistency learning methods with novel extensions of the mean-teacher (MT) model, which include a new auxiliary teacher, and the replacement of MT's mean square error (MSE) by a stricter confidence-weighted cross-entropy (Conf-CE) loss. The accurate prediction by this model allows us to use a challenging combination of network, input data and feature perturbations to improve the consistency learning generalisation, where the feature perturbations consist of a new adversarial perturbation. Results on public benchmarks show that our approach achieves remarkable improvements over the previous SOTA methods in the field.
****
Motion-Modulated Temporal Fragment Alignment Network for Few-Shot Action Recognition
Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9151-9160
While the majority of FSL models focus on image classification, the extension to action recognition is rather challenging due to the additional temporal dimension in videos. To address this issue, we propose an end-to-end Motion-modulated Temporal Fragment Alignment Network (MTFAN) by jointly exploring the task-specific motion modulation and the multi-level temporal fragment alignment for Few-Shot Action Recognition (FSAR). The proposed MTFAN model enjoys several merits. First, we design a motion modulator conditioned on the learned task-specific motion embeddings, which can activate the channels related to the task-shared motion patterns for each frame. Second, a segment attention mechanism is proposed to automatically discover the higher-level segments for multi-level temporal fragment alignment, which encompasses the frame-to-frame, segment-to-segment, and segment-to-frame alignments. To the best of our knowledge, this is the first work to exploit task-specific motion modulation for FSAR. Extensive experimental results on four standard benchmarks demonstrate that the proposed model performs favorably against the state-of-the-art FSAR methods.
****
Focal Sparse Convolutional Networks for 3D Object Detection
Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5428-5437
Non-uniformed 3D sparse data, e.g., point clouds or voxels in different spatial positions, make contribution to the task of 3D object detection in different ways. Existing basic components in sparse convolutional networks (Sparse CNNs) process all sparse data, regardless of regular or submanifold sparse convolution. In this paper, we introduce two new modules to enhance the capability of Sparse CNNs, both are based on making feature sparsity learnable with position-wise importance prediction. They are focal sparse convolution (Focals Conv) and its multi-modal variant of focal sparse convolution with fusion, or Focals Conv-F for short. The new modules can readily substitute their plain counterparts in existing Sparse CNNs and be jointly trained in an end-to-end fashion. For the first time, we show that spatially learnable sparsity in sparse convolution is essential for sophisticated 3D object detection. Extensive experiments on the KITTI, nuScenes and Waymo benchmarks validate the effectiveness of our approach. Without bells and whistles, our results outperform all existing single-model entries on the nuScenes test benchmark.

****
Masked Autoencoders Are Scalable Vision Learners
Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16000-16009
This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3x or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well: e.g., a vanilla ViT-Huge model achieves the best accuracy (87.8%) among methods that use only ImageNet-1K data. Transfer performance in downstream tasks outperforms supervised pre-training and shows promising scaling behavior.
****
Point-BERT: Pre-Training 3D Point Cloud Transformers With Masked Point Modeling
Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19313-19322
We present Point-BERT, a novel paradigm for learning Transformers to generalize the concept of BERT onto 3D point cloud. Following BERT, we devise a Masked Point Modeling (MPM) task to pre-train point cloud Transformers. Specifically, we first divide a point cloud into several local patches, and a point cloud Tokenizer is devised via a discrete Variational AutoEncoder (dVAE) to generate discrete point tokens containing meaningful local information. Then, we randomly mask some patches of input point clouds and feed them into the backbone Transformer. The pre-training objective is to recover the original point tokens at the masked locations under the supervision of point tokens obtained by the Tokenizer. Extensive experiments demonstrate that the proposed BERT-style pre-training strategy significantly improves the performance of standard point cloud Transformers. Equipped with our pre-training strategy, we show that a pure Transformer architecture attains 93.7% accuracy on ModelNet40 and 83.1% accuracy on the hardest setting of ScanObjectNN, surpassing carefully designed point cloud models with much fewer hand-made designs and human priors. We also demonstrate that the representations learned by Point-BERT transfer well to new tasks and domains, where our models largely advance the state-of-the-art of few-shot point cloud classification task.
****
Nested Collaborative Learning for Long-Tailed Visual Recognition
Jun Li, Zichang Tan, Jun Wan, Zhen Lei, Guodong Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6949-6958
The networks trained on the long-tailed dataset vary remarkably, despite the same training settings, which shows the great uncertainty in long-tailed learning. To alleviate the uncertainty, we propose a Nested Collaborative Learning (NCL), which tackles the problem by collaboratively learning multiple experts together. NCL consists of two core components, namely Nested Individual Learning (NIL) and Nested Balanced Online Distillation (NBOD), which focus on the individual supervised learning for each single expert and the knowledge transferring among multiple experts, respectively. To learn representations more thoroughly, both NIL and NBOD are formulated in a nested way, in which the learning is conducted on not just all categories from a full perspective but some hard categories from a partial perspective. Regarding the learning in the partial perspective, we specifically select the negative categories with high predicted scores as the hard categories by using a proposed Hard Category Mining (HCM). In the NCL, the learning

from two perspectives is nested, highly related and complementary, and helps the network to capture not only global and robust features but also meticulous distinguishing ability. Moreover, self-supervision is further utilized for feature enhancement. Extensive experiments manifest the superiority of our method with outperforming the state-of-the-art whether by using a single model or an ensemble.
****

Crowd Counting in the Frequency Domain
Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19618-19627
This paper investigates crowd counting in the frequency domain, which is a novel direction compared to the traditional view in the spatial domain. By transforming the density map into the frequency domain and using the nice properties of the characteristic function, we propose a novel method that is simple, effective, and efficient. The solid theoretical analysis ends up as an implementation-friendly loss function, which requires only standard tensor operations in the training process. We prove that our loss function is an upper bound of the pseudo sup norm metric between the ground truth and the prediction density map (over all of their sub-regions), and demonstrate its efficacy and efficiency versus other loss functions. The experimental results also show its competitiveness to the state-of-the-art on five benchmark data sets: ShanghaiTech A and B, UCF-QNRF, JHU++, and NWPU.
****

Restormer: Efficient Transformer for High-Resolution Image Restoration
Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5728-5739
Since convolutional neural networks (CNNs) perform well at learning generalizable image priors from large-scale data, these models have been extensively applied to image restoration and related tasks. Recently, another class of neural architectures, Transformers, have shown significant performance gains on natural language and high-level vision tasks. While the Transformer model mitigates the shortcomings of CNNs (i.e., limited receptive field and inadaptability to input content), its computational complexity grows quadratically with the spatial resolution, therefore making it infeasible to apply to most image restoration tasks involving high-resolution images. In this work, we propose an efficient Transformer model by making several key designs in the building blocks (multi-head attention and feed-forward network) such that it can capture long-range pixel interactions, while still remaining applicable to large images. Our model, named Restoration Transformer (Restormer), achieves state-of-the-art results on several image restoration tasks, including image deraining, single-image motion deblurring, defocus deblurring (single-image and dual-pixel data), and image denoising (Gaussian grayscale/color denoising, and real image denoising). The source code and pre-trained models are available at https://github.com/swz30/Restormer.
****

STRPM: A Spatiotemporal Residual Predictive Model for High-Resolution Video Prediction
Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Wen Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13946-13955
Although many video prediction methods have obtained good performance in low-resolution (64 128) videos, predictive models for high-resolution (512 4K) videos have not been fully explored yet, which are more meaningful due to the increasing demand for high-quality videos. Compared with low-resolution videos, high-resolution videos contain richer appearance (spatial) information and more complex motion (temporal) information. In this paper, we propose a Spatiotemporal Residual Predictive Model (STRPM) for high-resolution video prediction. On the one hand, we propose a Spatiotemporal Encoding-Decoding Scheme to preserve more spatiotemporal information for high-resolution videos. In this way, the appearance details for each frame can be greatly preserved. On the other hand, we design a Residu

al Predictive Memory (RPM) which focuses on modeling the spatiotemporal residual features (STRF) between previous and future frames instead of the whole frame, which can greatly help capture the complex motion information in high-resolution videos. In addition, the proposed RPM can supervise the spatial encoder and temporal encoder to extract different features in the spatial domain and the temporal domain, respectively. Moreover, the proposed model is trained using generative adversarial networks (GANs) with a learned perceptual loss (LP-loss) to improve the perceptual quality of the predictions. Experimental results show that STRPM can generate more satisfactory results compared with various existing methods.
****

Learning From Untrimmed Videos: Self-Supervised Video Representation Learning With Hierarchical Consistency
Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yi Xu, Xiang Wang, Mingqian Tang, Changxin Gao, Rong Jin, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13821-13831
Natural videos provide rich visual contents for self-supervised learning. Yet most existing approaches for learning spatio-temporal representations rely on manually trimmed videos, leading to limited diversity in visual patterns and limited performance gain. In this work, we aim to learn representations by leveraging more abundant information in untrimmed videos. To this end, we propose to learn a hierarchy of consistencies in videos, i.e., visual consistency and topical consistency, corresponding respectively to clip pairs that tend to be visually similar when separated by a short time span and share similar topics when separated by a long time span. Specifically, a hierarchical consistency learning framework HiCo is presented, where the visually consistent pairs are encouraged to have the same representation through contrastive learning, while the topically consistent pairs are coupled through a topical classifier that distinguishes whether they are topicrelated. Further, we impose a gradual sampling algorithm for proposed hierarchical consistency learning, and demonstrate its theoretical superiority. Empirically, we show that not only HiCo can generate stronger representations on untrimmed videos, it also improves the representation quality when applied to trimmed videos. This is in contrast to standard contrastive learning that fails to learn appropriate representations from untrimmed videos.
****

Aladdin: Joint Atlas Building and Diffeomorphic Registration Learning With Pairwise Alignment
Zhipeng Ding, Marc Niethammer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20784-20793
Atlas building and image registration are important tasks for medical image analysis. Once one or multiple atlases from an image population have been constructed, commonly (1) images are warped into an atlas space to study intra-subject or inter-subject variations or (2) a possibly probabilistic atlas is warped into image space to assign anatomical labels. Atlas estimation and nonparametric transformations are computationally expensive as they usually require numerical optimization. Additionally, previous approaches for atlas building often define similarity measures between a fuzzy atlas and each individual image, which may cause alignment difficulties because a fuzzy atlas does not exhibit clear anatomical structures in contrast to the individual images. This work explores using a convolutional neural network (CNN) to jointly predict the atlas and a stationary velocity field (SVF) parameterization for diffeomorphic image registration with respect to the atlas. Our approach does not require affine pre-registrations and utilizes pairwise image alignment losses to increase registration accuracy. We evaluate our model on 3D knee magnetic resonance images (MRI) from the OAI-ZIB dataset. Our results show that the proposed framework achieves better performance than other state-of-the-art image registration algorithms, allows for end-to-end training, and for fast inference at test time.
****

IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation
Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, Jie Yang; Proceedings of the IEEE/CVF Conference on Computer Vis

ion and Pattern Recognition (CVPR), 2022, pp. 1969-1978
Prevailing video frame interpolation algorithms, that generate the intermediate frames from consecutive inputs, typically rely on complex model architectures with heavy parameters or large delay, hindering them from diverse real-time applications. In this work, we devise an efficient encoder-decoder based network, termed IFRNet, for fast intermediate frame synthesizing. It first extracts pyramid features from given inputs, and then refines the bilateral intermediate flow fields together with a powerful intermediate feature until generating the desired output. The gradually refined intermediate feature can not only facilitate intermediate flow estimation, but also compensate for contextual details, making IFRNet do not need additional synthesis or refinement module. To fully release its potential, we further propose a novel task-oriented optical flow distillation loss to focus on learning the useful teacher knowledge towards frame synthesizing. Meanwhile, a new geometry consistency regularization term is imposed on the gradually refined intermediate features to keep better structure layout. Experiments on various benchmarks demonstrate the excellent performance and fast inference speed of proposed approaches. Code is available at https://github.com/ltkong218/IFRNet.
****

Large Loss Matters in Weakly Supervised Multi-Label Classification
Youngwook Kim, Jae Myung Kim, Zeynep Akata, Jungwoo Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14156-14165
Weakly supervised multi-label classification (WSML) task, which is to learn a multi-label classification using partially observed labels per image, is becoming increasingly important due to its huge annotation cost. In this work, we first regard unobserved labels as negative labels, casting the WSML task into noisy multi-label classification. From this point of view, we empirically observe that memorization effect, which was first discovered in a noisy multi-class setting, also occurs in a multi-label setting. That is, the model first learns the representation of clean labels, and then starts memorizing noisy labels. Based on this finding, we propose novel methods for WSML which reject or correct the large loss samples to prevent model from memorizing the noisy label. Without heavy and complex components, our proposed methods outperform previous state-of-the-art WSML methods on several partial label settings including Pascal VOC 2012, MS COCO, NUSWIDE, CUB, and OpenImages V3 datasets. Various analysis also show that our methodology actually works well, validating that treating large loss properly matters in a weakly supervised multi-label classification. Our code is available at https://github.com/snucml/LargeLossMatters.
****

Toward Practical Monocular Indoor Depth Estimation
Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, Shuochen Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3814-3824
The majority of prior monocular depth estimation methods without groundtruth depth guidance focus on driving scenarios. We show that such methods generalize poorly to unseen complex indoor scenes, where objects are cluttered and arbitrarily arranged in the near field. To obtain more robustness, we propose a structure distillation approach to learn knacks from an off-the-shelf relative depth estimator that produces structured but metric-agnostic depth. By combining structure distillation with a branch that learns metrics from left-right consistency, we attain structured and metric depth for generic indoor scenes and make inferences in real-time. To facilitate learning and evaluation, we collect SimSIN, a dataset from simulation with thousands of environments, and UniSIN, a dataset that contains about 500 real scan sequences of generic indoor environments. We experiment in both sim-to-real and real-to-real settings, and show improvements, as well as in downstream applications using our depth maps. This work provides a full study, covering methods, data, and applications aspects.
****

Attention Concatenation Volume for Accurate and Efficient Stereo Matching

Gangwei Xu, Junda Cheng, Peng Guo, Xin Yang; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12981-12990
Stereo matching is a fundamental building block for many vision and robotics app lications. An informative and concise cost volume representation is vital for st ereo matching of high accuracy and efficiency. In this paper, we present a novel cost volume construction method which generates attention weights from correlat ion clues to suppress redundant information and enhance matching-related informa tion in the concatenation volume. To generate reliable attention weights, we pro pose multi-level adaptive patch matching to improve the distinctiveness of the m atching cost at different disparities even for textureless regions. The proposed cost volume is named attention concatenation volume (ACV) which can be seamless ly embedded into most stereo matching networks, the resulting networks can use a more lightweight aggregation network and meanwhile achieve higher accuracy, e.g . using only 1/25 parameters of the aggregation network can achieve higher accur acy for GwcNet. Furthermore, we design a highly accurate network (ACVNet) based on our ACV, which achieves state-of-the-art performance on several benchmarks.
****

Learning Distinctive Margin Toward Active Domain Adaptation
Ming Xie, Yuxi Li, Yabiao Wang, Zekun Luo, Zhenye Gan, Zhongyi Sun, Mingmin Chi, Chengjie Wang, Pei Wang; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2022, pp. 7993-8002
Despite plenty of efforts focusing on improving the domain adaptation ability (D A) under unsupervised or few-shot semi-supervised settings, recently the solutio n of active learning started to attract more attention due to its suitability in transferring model in a more practical way with limited annotation resource on target data. Nevertheless, most active learning methods are not inherently desig ned to handle domain gap between data distribution, on the other hand, some acti ve domain adaptation methods (ADA) usually requires complicated query functions, which is vulnerable to overfitting. In this work, we propose a concise but effe ctive ADA method called Select-by-Distinctive-Margin (SDM), which consists of a maximum margin loss and a margin sampling algorithm for data selection. We provi de theoretical analysis to show that SDM works like a Support Vector Machine, st oring hard examples around decision boundaries and exploiting them to find infor mative and transferable data. In addition, we propose two variants of our method , one is designed to adaptively adjust the gradient from margin loss, the other boosts the selectivity of margin sampling by taking the gradient direction into account. We benchmark SDM with standard active learning setting, demonstrating o ur algorithm achieves competitive results with good data scalability.
****

Zero-Query Transfer Attacks on Context-Aware Object Detectors
Zikui Cai, Shantanu Rane, Alejandro E. Brito, Chengyu Song, Srikanth V. Krishnam urthy, Amit K. Roy-Chowdhury, M. Salman Asif; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15024-15034
Adversarial attacks perturb images such that a deep neural network produces inco rrect classification results. A promising approach to defend against adversarial attacks on natural multi-object scenes is to impose a context-consistency check , wherein, if the detected objects are not consistent with an appropriately defi ned context, then an attack is suspected. Stronger attacks are needed to fool su ch context-aware detectors. We present the first approach for generating context -consistent adversarial attacks that can evade the context-consistency check of black-box object detectors operating on complex, natural scenes. Unlike many bla ck-box attacks that perform repeated attempts and open themselves to detection, we assume a "zero-query" setting, where the attacker has no knowledge of the cla ssification decisions of the victim system. First, we derive multiple attack pla ns that assign incorrect labels to victim objects in a context-consistent manner . Then we design and use a novel data structure that we call the perturbation su ccess probability matrix, which enables us to filter the attack plans and choose the one most likely to succeed. This final attack plan is implemented using a p erturbation-bounded adversarial attack algorithm. We compare our zero-query atta ck against a few-query scheme that repeatedly checks if the victim system is foo

led. We also compare against state-of-the-art context-agnostic attacks. Against a context-aware defense, the fooling rate of our zero-query approach is significantly higher than context-agnostic approaches and higher than that achievable with up to three rounds of the few-query scheme.
****

Neural Inertial Localization
Sachini Herath, David Caruso, Chen Liu, Yufan Chen, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6604-6613
This paper proposes the inertial localization problem, the task of estimating the absolute location from a sequence of inertial sensor measurements. This is an exciting and unexplored area of indoor localization research, where we present a rich dataset with 53 hours of inertial sensor data and the associated ground truth locations. We developed a solution, dubbed neural inertial localization (NILoc) which 1) uses a neural inertial navigation technique to turn inertial sensor history to a sequence of velocity vectors; then 2) employs a transformer-based neural architecture to find the device location from the sequence of velocity estimates. We only use an IMU sensor, which is energy efficient and privacy-preserving compared to WiFi, cameras, and other data sources. Our approach is significantly faster and achieves competitive results even compared with state-of-the-art methods that require a floorplan and run 20 to 30 times slower. We share our code, model and data at https://sachini.github.io/niloc.
****

Speed Up Object Detection on Gigapixel-Level Images With Patch Arrangement
Jiahao Fan, Huabin Liu, Wenjie Yang, John See, Aixin Zhang, Weiyao Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4653-4661
With the appearance of super high-resolution (e.g., gigapixel-level) images, performing efficient object detection on such images becomes an important issue. Most existing works for efficient object detection on high-resolution images focus on generating local patches where objects may exist, and then every patch is detected independently. However, when the image resolution reaches gigapixel-level, they will suffer from a huge time cost for detecting numerous patches. Different from them, we devise a novel patch arrangement framework for fast object detection on gigapixel-level images. Under this framework, a Patch Arrangement Network (PAN) is proposed to accelerate the detection by determining which patches could be packed together into a compact canvas. Specifically, PAN consists of (1) a Patch Filter Module (PFM) (2) a Patch Packing Module (PPM). PFM filters patch candidates by learning to select patches between two granularities. Subsequently, from the remaining patches, PPM determines how to pack these patches together into a smaller number of canvases. Meanwhile, it generates an ideal layout of patches on canvas. These canvases are fed to the detector to get final results. Experiments show that our method could improve the inference speed on gigapixel-level images by 5 times while maintaining great performance.
****

Finding Fallen Objects via Asynchronous Audio-Visual Integration
Chuang Gan, Yi Gu, Siyuan Zhou, Jeremy Schwartz, Seth Alter, James Traer, Dan Gutfreund, Joshua B. Tenenbaum, Josh H. McDermott, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10523-10533
The way an object looks and sounds provide complementary reflections of its physical properties. In many settings cues from vision and audition arrive asynchronously but must be integrated, as when we hear an object dropped on the floor and then must find it. In this paper, we introduce a setting in which to study multi-modal object localization in 3D virtual environments. An object is dropped somewhere in a room. An embodied robot agent, equipped with a camera and microphone, must determine what object has been dropped -- and where -- by combining audio and visual signals with knowledge of the underlying physics. To study this problem, we have generated a large-scale dataset -- the Fallen Objects dataset -- that includes 8000 instances of 30 physical object categories in 64 rooms. The dat

aset uses the ThreeDWorld Platform that can simulate physics-based impact sounds and complex physical interactions between objects in a photorealistic setting. As a first step toward addressing this challenge, we develop a set of embodied agent baselines, based on imitation learning, reinforcement learning, and modular planning, and perform an in-depth analysis of the challenge of this new task.
****

Learning sRGB-to-Raw-RGB De-Rendering With Content-Aware Metadata
Seonghyeon Nam, Abhijith Punnappurath, Marcus A. Brubaker, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17704-17713
Most camera images are rendered and saved in the standard RGB (sRGB) format by the camera's hardware. Due to the in-camera photo-finishing routines, nonlinear sRGB images are undesirable for computer vision tasks that assume a direct relationship between pixel values and scene radiance. For such applications, linear raw-RGB sensor images are preferred. Saving images in their raw-RGB format is still uncommon due to the large storage requirement and lack of support by many imaging applications. Several "raw reconstruction" methods have been proposed that utilize specialized metadata sampled from the raw-RGB image at capture time and embedded in the sRGB image. This metadata is used to parameterize a mapping function to de-render the sRGB image back to its original raw-RGB format when needed. Existing raw reconstruction methods rely on simple sampling strategies and global mapping to perform the de-rendering. This paper shows how to improve the de-rendering results by jointly learning sampling and reconstruction. Our experiments show that our learned sampling can adapt to the image content to produce better raw reconstructions than existing methods. We also describe an online fine-tuning strategy for the reconstruction network to improve results further.
****

GraftNet: Towards Domain Generalized Stereo Matching With a Broad-Spectrum and Task-Oriented Feature
Biyang Liu, Huimin Yu, Guodong Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13012-13021
Although supervised deep stereo matching networks have made impressive achievements, the poor generalization ability caused by the domain gap prevents them from being applied to real-life scenarios. In this paper, we propose to leverage the feature of a model trained on large-scale datasets to deal with the domain shift since it has seen various styles of images. With the cosine similarity based cost volume as a bridge, the feature will be grafted to an ordinary cost aggregation module. Despite the broad-spectrum representation, such a low-level feature contains much general information which is not aimed at stereo matching. To recover more task-specific information, the grafted feature is further input into a shallow network to be transformed before calculating the cost. Extensive experiments show that the model generalization ability can be improved significantly with this broad-spectrum and task-oriented feature. Specifically, based on two well-known architectures PSMNet and GANet, our methods are superior to other robust algorithms when transferring from SceneFlow to KITTI 2015, KITTI 2012, and Middlebury. Code is available at https://github.com/SpadeLiu/Graft-PSMNet.
****

Towards Total Recall in Industrial Anomaly Detection
Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, Peter Gehler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14318-14328
Being able to spot defective parts is a critical component in large-scale industrial manufacturing. A particular challenge that we address in this work is the cold-start problem: fit a model using nominal (non-defective) example images only. While handcrafted solutions per class are possible, the goal is to build systems that work well simultaneously on many different tasks automatically. The best peforming approaches combine embeddings from ImageNet models with an outlier detection model. In this paper, we extend on this line of work and propose PatchCore, which uses a maximally representative memory bank of nominal patch-features. PatchCore offers competitive inference times while achieving state-of-the-art p

erformance for both detection and localization. On the challenging, widely used MVTec AD benchmark PatchCore achieves an image-level anomaly detection AUROC score of up to 99.6%, more than halving the error compared to the next best competitor. We further report competitive results on two additional datasets and also find competitive results in the few samples regime. Code: github.com/amazon-research/patchcore-inspection

****

DTA: Physical Camouflage Attacks Using Differentiable Transformation Network

Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, Howon Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15305-15314

To perform adversarial attacks in the physical world, many studies have proposed adversarial camouflage, a method to hide a target object by applying camouflage patterns on 3D object surfaces. For obtaining optimal physical adversarial camouflage, previous studies have utilized the so-called neural renderer, as it supports differentiability. However, existing neural renderers cannot fully represent various real-world transformations due to a lack of control of scene parameters compared to the legacy photo-realistic renderers. In this paper, we propose the Differentiable Transformation Attack (DTA), a framework for generating a robust physical adversarial pattern on a target object to camouflage it against object detection models with a wide range of transformations. It utilizes our novel Differentiable Transformation Network (DTN), which learns the expected transformation of a rendered object when the texture is changed while preserving the original properties of the target object. Using our attack framework, an adversary can gain both the advantages of the legacy photo-realistic renderers including various physical-world transformations and the benefit of white-box access by offering differentiability. Our experiments show that our camouflaged 3D vehicles can successfully evade state-of-the-art object detection models in the photo-realistic environment (i.e., CARLA on Unreal Engine). Furthermore, our demonstration on a scaled Tesla Model 3 proves the applicability and transferability of our method to the real world.

****

Neural Recognition of Dashed Curves With Gestalt Law of Continuity

Hanyuan Liu, Chengze Li, Xueting Liu, Tien-Tsin Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1373-1382

Dashed curve is a frequently used curve form and is widely used in various drawing and illustration applications. While humans can intuitively recognize dashed curves from disjoint curve segments based on the law of continuity in Gestalt psychology, it is extremely difficult for computers to model the Gestalt law of continuity and recognize the dashed curves since high-level semantic understanding is needed for this task. The various appearances and styles of the dashed curves posed on a potentially noisy background further complicate the task. In this paper, we propose an innovative Transformer-based framework to recognize dashed curves based on both high-level features and low-level clues. The framework manages to learn the computational analogy of the Gestalt Law in various domains to locate and extract instances of dashed curves in both raster and vector representations. Qualitative and quantitative evaluations demonstrate the efficiency and robustness of our framework over all existing solutions.

****

Semi-Supervised Object Detection via Multi-Instance Alignment With Global Class Prototypes

Aoxue Li, Peng Yuan, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9809-9818

Semi-Supervised object detection (SSOD) aims to improve the generalization ability of object detectors with large-scale unlabeled images. Current pseudo-labeling-based SSOD methods individually learn from labeled data and unlabeled data, without considering the relation between them. To make full use of labeled data, we propose a Multi-instance Alignment model which enhances the prediction consist

ency based on Global Class Prototypes (MA-GCP). Specifically, we impose the cons istency between pseudo ground-truths and their high-IoU candidates by minimizing the cross-entropy loss of their class distributions computed based on global cl ass prototypes. These global class prototypes are estimated with the whole label ed dataset via the exponential moving average algorithm. To evaluate the propose d MA-GCP model, we integrate it into the state-of-the-art SSOD framework and exp eriments on two benchmark datasets demonstrate the effectiveness of our MA-GCP a pproach.

****

HODOR: High-Level Object Descriptors for Object Re-Segmentation in Video Learned From Static Images

Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, Bastian Leibe; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 3022-3031

Existing state-of-the-art methods for Video Object Segmentation (VOS) learn low- level pixel-to-pixel correspondences between frames to propagate object masks ac ross video. This requires a large amount of densely annotated video data, which is costly to annotate, and largely redundant since frames within a video are hig hly correlated. In light of this, we propose HODOR: a novel method that tackles VOS by effectively leveraging annotated static images for understanding object a ppearance and scene context. We encode object instances and scene information fr om an image frame into robust high-level descriptors which can then be used to r e-segment those objects in different frames. As a result, HODOR achieves state-o f-the-art performance on the DAVIS and YouTube-VOS benchmarks compared to existi ng methods trained without video annotations. Without any architectural modifica tion, HODOR can also learn from video context around single annotated video fram es by utilizing cyclic consistency, whereas other methods rely on dense, tempora lly consistent annotations.

****

Point Cloud Color Constancy

Xiaoyan Xing, Yanlin Qian, Sibo Feng, Yuhan Dong, Ji■í Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp . 19750-19759

In this paper, we present Point Cloud Color Constancy, in short PCCC, an illumin ation chromaticity estimation algorithm exploiting a point cloud. We leverage th e depth information captured by the time-of-flight (ToF) sensor mounted rigidly with the RGB sensor, and form a 6D cloud where each point contains the coordinat es and RGB intensities, noted as (x,y,z,r,g,b). PCCC applies the PointNet archit ecture to the color constancy problem, deriving the illumination vector point-wi se and then making a global decision about the global illumination chromaticity. On two popular RGB-D datasets, which we extend with illumination information, a s well as on a novel benchmark, PCCC obtains lower error than the state-of-the-a rt algorithms. Our method is simple and fast, requiring merely 16*16-size input and reaching speed over 500 fps, including the cost of building the point cloud and net inference.

****

VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning

Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, Zeynep Akata; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 9316-9325

Human-annotated attributes serve as powerful semantic embeddings in zero-shot le arning. However, their annotation process is labor-intensive and needs expert su pervision. Current unsupervised semantic embeddings, i.e., word embeddings, enab le knowledge transfer between classes. However, word embeddings do not always re flect visual similarities and result in inferior zero-shot performance. We propo se to discover semantic embeddings containing discriminative visual properties f or zero-shot learning, without requiring any human annotation. Our model visuall y divides a set of images from seen classes into clusters of local image regions according to their visual similarity, and further imposes their class discrimin ation and semantic relatedness. To associate these clusters with previously unse

en classes, we use external knowledge, e.g., word embeddings and propose a novel class relation discovery module. Through quantitative and qualitative evaluation, we demonstrate that our model discovers semantic embeddings that model the visual properties of both seen and unseen classes. Furthermore, we demonstrate on three benchmarks that our visually-grounded semantic embeddings further improve performance over word embeddings across various ZSL models by a large margin. Code is available at https://github.com/wenjiaXu/VGSE

****

Catching Both Gray and Black Swans: Open-Set Supervised Anomaly Detection
Choubo Ding, Guansong Pang, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7388-7398

Despite most existing anomaly detection studies assume the availability of normal training samples only, a few labeled anomaly examples are often available in many real-world applications, such as defect samples identified during random quality inspection, lesion images confirmed by radiologists in daily medical screening, etc. These anomaly examples provide valuable knowledge about the application-specific abnormality, enabling significantly improved detection of similar anomalies in some recent models. However, those anomalies seen during training often do not illustrate every possible class of anomaly, rendering these models ineffective in generalizing to unseen anomaly classes. This paper tackles open-set supervised anomaly detection, in which we learn detection models using the anomaly examples with the objective to detect both seen anomalies ('gray swans') and unseen anomalies ('black swans'). We propose a novel approach that learns disentangled representations of abnormalities illustrated by seen anomalies, pseudo anomalies, and latent residual anomalies (i.e., samples that have unusual residuals compared to the normal data in a latent space), with the last two abnormalities designed to detect unseen anomalies. Extensive experiments on nine real-world anomaly detection datasets show superior performance of our model in detecting seen and unseen anomalies under diverse settings. Code and data are available at: https://github.com/choubo/DRA.

****

MLSLT: Towards Multilingual Sign Language Translation
Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, Xiaofei He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5109-5119

Most of the research to date focuses on bilingual sign language translation (BSLT). However, such models are inefficient in building multilingual sign language translation systems. To solve this problem, we introduce the multilingual sign language translation (MSLT) task. It aims to use a single model to complete the translation between multiple sign languages and spoken languages. Then, we propose MLSLT, the first MSLT model, which contains two novel dynamic routing mechanisms for controlling the degree of parameter sharing between different languages. Intra-layer language-specific routing controls the proportion of data flowing through shared parameters and language-specific parameters from the token level through a soft gate within the layer, and inter-layer language-specific routing controls and learns the data flow path of different languages at the language level through a soft gate between layers. In order to evaluate the performance of MLSLT, we collect the first publicly available multilingual sign language understanding dataset, Spreadthesign-Ten (SP-10), which contains up to 100 language pairs, e.g., CSL->en, GSG->zh. Experimental results show that the average performance of MLSLT outperforms the baseline MSLT model and the combination of multiple BSLT models in many cases. In addition, we also explore zero-shot translation in sign language and find that our model can achieve comparable performance to the supervised BSLT model on some language pairs. Dataset and more details are at https://mlslt.github.io/.

****

Towards an End-to-End Framework for Flow-Guided Video Inpainting
Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17562-17571

Optical flow, which captures motion information across frames, is exploited in r ecent video inpainting methods through propagating pixels along its trajectories . However, the hand-crafted flow-based processes in these methods are applied se parately to form the whole inpainting pipeline. Thus, they are less efficient an d rely heavily on the intermediate results from earlier stages. In this paper, w e propose an End-to-End framework for Flow-Guided Video Inpainting through elabo rately designed three trainable modules, namely, flow completion, feature propag ation, and content hallucination modules. The three modules correspond with the three stages of previous flow-based methods but can be jointly optimized, leadin g to a more efficient and effective inpainting process. Experimental results dem onstrate that our proposed method outperforms state-of-the-art methods both qual itatively and quantitatively and shows promising efficiency. The code is availab le at https://github.com/MCG-NKU/E2FGVI.

****

Contrastive Test-Time Adaptation

Dian Chen, Dequan Wang, Trevor Darrell, Sayna Ebrahimi; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 295-305

Test-time adaptation is a special setting of unsupervised domain adaptation wher e a trained model on the source domain has to adapt to the target domain without accessing source data. We propose a novel way to leverage self-supervised contr astive learning to facilitate target feature learning, along with an online pseu do labeling scheme with refinement that significantly denoises pseudo labels. Th e contrastive learning task is applied jointly with pseudo labeling, contrasting positive and negative pairs constructed similarly as MoCo but with source-initi alized encoder, and excluding same-class negative pairs indicated by pseudo labe ls. Meanwhile, we produce pseudo labels online and refine them via soft voting a mong their nearest neighbors in the target feature space, enabled by maintaining a memory queue. Our method, AdaContrast, achieves state-of-the-art performance on major benchmarks while having several desirable properties compared to existi ng works, including memory efficiency, insensitivity to hyper-parameters, and be tter model calibration. Code is released at https://github.com/DianCh/AdaContras t.

****

Multimodal Colored Point Cloud to Image Alignment

Noam Rotstein, Amit Bracha, Ron Kimmel; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6656-6666

Reconstruction of geometric structures from images using supervised learning suf fers from limited available amount of accurate data. One type of such data is ac curate real-world RGB-D images. A major challenge in acquiring such ground truth data is the accurate alignment between RGB images and the point cloud measured by a depth scanner. To overcome this difficulty, we consider a differential opti mization method that aligns a colored point cloud with a given color image throu gh iterative geometric and color matching. In the proposed framework, the optimi zation minimizes the photometric difference between the colors of the point clou d and the corresponding colors of the image pixels. Unlike other methods that tr y to reduce this photometric error, we analyze the computation of the gradient o n the image plane and propose a different direct scheme. We assume that the colo rs produced by the geometric scanner camera and the color camera sensor are diff erent and therefore characterized by different chromatic acquisition properties. Under these multimodal conditions, we find the transformation between the camer a image and the point cloud colors. We alternately optimize for aligning the pos ition of the point cloud and matching the different color spaces. The alignments produced by the proposed method are demonstrated on both synthetic data with qu antitative evaluation and real scenes with qualitative results.

****

MotionAug: Augmentation With Physical Correction for Human Motion Prediction

Takahiro Maeda, Norimichi Ukita; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2022, pp. 6427-6436

This paper presents a motion data augmentation scheme incorporating motion synth

esis encouraging diversity and motion correction imposing physical plausibility. This motion synthesis consists of our modified Variational AutoEncoder (VAE) and Inverse Kinematics (IK). In this VAE, our proposed sampling-near-samples method generates various valid motions even with insufficient training motion data. Our IK-based motion synthesis method allows us to generate a variety of motions semi-automatically. Since these two schemes generate unrealistic artifacts in the synthesized motions, our motion correction rectifies them. This motion correction scheme consists of imitation learning with physics simulation and subsequent motion debiasing. For this imitation learning, we propose the PD-residual force that significantly accelerates the training process. Furthermore, our motion debiasing successfully offsets the motion bias induced by imitation learning to maximize the effect of augmentation. As a result, our method outperforms previous noise-based motion augmentation methods by a large margin on both Recurrent Neural Network-based and Graph Convolutional Network-based human motion prediction models. The code is available at https://github.com/meaten/MotionAug.

****

Active Teacher for Semi-Supervised Object Detection

Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14482-14491

In this paper, we study teacher-student learning from the perspective of data initialization and propose a novel algorithm called Active Teacher for semi-supervised object detection (SSOD). Active Teacher extends the teacher-student framework to an iterative version, where the label set is partially initialized and gradually augmented by evaluating three key factors of unlabeled examples, including difficulty, information and diversity. With this design, Active Teacher can maximize the effect of limited label information while improving the quality of pseudo-labels. To validate our approach, we conduct extensive experiments on the MS-COCO benchmark and compare Active Teacher with a set of recently proposed SSOD methods. The experimental results not only validate the superior performance gain of Active Teacher over the compared methods, but also show that it enables the baseline network, ie, Faster-RCNN, to achieve 100% supervised performance with much less label expenditure, ie 40% labeled examples on MS-COCO. More importantly, we believe that the experimental analyses in this paper can provide useful empirical knowledge for data annotation in practical applications.

****

CrossLoc: Scalable Aerial Localization Assisted by Multimodal Synthetic Data

Qi Yan, Jianhao Zheng, Simon Reding, Shanci Li, Iordan Doytchinov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17358-17368

We present a visual localization system that learns to estimate camera poses in the real world with the help of synthetic data. Despite significant progress in recent years, most learning-based approaches to visual localization target at a single domain and require a dense database of geo-tagged images to function well. To mitigate the data scarcity issue and improve the scalability of the neural localization models, we introduce TOPO-DataGen, a versatile synthetic data generation tool that traverses smoothly between the real and virtual world, hinged on the geographic camera viewpoint. New large-scale sim-to-real benchmark datasets are proposed to showcase and evaluate the utility of the said synthetic data. Our experiments reveal that synthetic data generically enhances the neural network performance on real data. Furthermore, we introduce CrossLoc, a cross-modal visual representation learning approach to pose estimation that makes full use of the scene coordinate ground truth via self-supervision. Without any extra data, CrossLoc significantly outperforms the state-of-the-art methods and achieves substantially higher real-data sample efficiency. Our code and datasets are all available at crossloc.github.io.

****

Audio-Adaptive Activity Recognition Across Video Domains

Yunhua Zhang, Hazel Doughty, Ling Shao, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13

791-13800
This paper strives for activity recognition under domain shift, for example caus ed by change of scenery or camera viewpoint. The leading approaches reduce the s hift in activity appearance by adversarial training and self-supervised learning . Different from these vision-focused works we leverage activity sounds for doma in adaptation as they have less variance across domains and can reliably indicat e which activities are not happening. We propose an audio-adaptive encoder and a ssociated learning methods that discriminatively adjust the visual feature repre sentation as well as addressing shifts in the semantic distribution. To further eliminate domain-specific features and include domain-invariant activity sounds for recognition, an audio-infused recognizer is proposed, which effectively mode ls the cross-modal interaction across domains. We also introduce the new task of actor shift, with a corresponding audio-visual dataset, to challenge our method with situations where the activity appearance changes dramatically. Experiments on this dataset, EPIC-Kitchens and CharadesEgo show the effectiveness of our ap proach.
****

Collaborative Learning for Hand and Object Reconstruction With Attention-Guided Graph Convolution
Tze Ho Elden Tse, Kwang In Kim, Ales■ Leonardis, Hyung Jin Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 1664-1674
Estimating the pose and shape of hands and objects under interaction finds numer ous applications including augmented and virtual reality. Existing approaches fo r hand and object reconstruction require explicitly defined physical constraints and known objects, which limits its application domains. Our algorithm is agnos tic to object models, and it learns the physical rules governing hand-object int eraction.This requires automatically inferring the shapes and physical interacti on of hands and (potentially unknown) objects. We seek to approach this challeng ing problem by proposing a collaborative learning strategy where two-branches of deep networks are learning from each other. Specifically, we transfer hand mesh information to the object branch and vice versa for the hand branch. The result ing optimisation (training) problem can be unstable, and we address this via two strategies: (i) attention-guided graph convolution which helps identify and foc us on mutual occlusion and (ii) unsupervised associative loss which facilitates the transfer of information between the branches. Experiments using four widely- used benchmarks show that our framework achieves beyond state-of-the-art accurac y in 3D pose estimation, as well as recovers dense 3D hand and object shapes. Ea ch technical component above contributes meaningfully in the ablation study.
****

On Learning Contrastive Representations for Learning With Noisy Labels
Li Yi, Sheng Liu, Qi She, A. Ian McLeod, Boyu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16682-16 691
Deep neural networks are able to memorize noisy labels easily with a softmax cro ss entropy (CE) loss. Previous studies attempted to address this issue focus on incorporating a noise-robust loss function to the CE loss. However, the memoriza tion issue is alleviated but still remains due to the non-robust CE loss. To add ress this issue, we focus on learning robust contrastive representations of data on which the classifier is hard to memorize the label noise under the CE loss. We propose a novel contrastive regularization function to learn such representat ions over noisy data where the label noise does not dominate the representation learning. By theoretically investigating the representations induced by the prop osed regularization function, we reveal that the learned representations keep in formation related to true labels and discard information related to corrupted la bels from images. Moreover, our theoretical results also indicate that the learn ed representations are robust to the label noise. Experiments on benchmark datas ets demonstrate that the efficacy of our method.
****

Unsupervised Deraining: Where Contrastive Learning Meets Self-Similarity

Yuntong Ye, Changfeng Yu, Yi Chang, Lin Zhu, Xi-Le Zhao, Luxin Yan, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5821-5830

Image deraining is a typical low-level image restoration task, which aims at decomposing the rainy image into two distinguishable layers: the clean image layer and the rain layer. Most of the existing learning-based deraining methods are supervisedly trained on synthetic rainy-clean pairs. The domain gap between the synthetic and real rains makes them less generalized to different real rainy scenes. Moreover, the existing methods mainly utilize the property of the two layers independently, while few of them have considered the mutually exclusive relationship between the two layers. In this work, we propose a novel non-local contrastive learning (NLCL) method for unsupervised image deraining. Consequently, we not only utilize the intrinsic self-similarity property within samples but also the mutually exclusive property between the two layers, so as to better differ the rain layer from the clean image. Specifically, the non-local self-similarity image layer patches as the positives are pulled together and similar rain layer patches as the negatives are pushed away. Thus the similar positive/negative samples that are close in the original space benefit us to enrich more discriminative representation. Apart from the self-similarity sampling strategy, we analyze how to choose an appropriate feature encoder in NLCL. Extensive experiments on different real rainy datasets demonstrate that the proposed method obtains state-of-the-art performance in real deraining.

****

Modeling Indirect Illumination for Inverse Rendering

Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18643-18652

Recent advances in implicit neural representations and differentiable rendering make it possible to simultaneously recover the geometry and materials of an object from multi-view RGB images captured under unknown static illumination. Despite the promising results achieved, indirect illumination is rarely modeled in previous methods, as it requires expensive recursive path tracing which makes the inverse rendering computationally intractable. In this paper, we propose a novel approach to efficiently recovering spatially-varying indirect illumination. The key insight is that indirect illumination can be conveniently derived from the neural radiance field learned from input images instead of being estimated jointly with direct illumination and materials. By properly modeling the indirect illumination and visibility of direct illumination, interreflection- and shadow-free albedo can be recovered. The experiments on both synthetic and real data demonstrate the superior performance of our approach compared to previous work and its capability to synthesize realistic renderings under novel viewpoints and illumination. Our code and data are available at https://zju3dv.github.io/invrender/.

****

BACON: Band-Limited Coordinate Networks for Multiscale Scene Representation

David B. Lindell, Dave Van Veen, Jeong Joon Park, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16252-16262

Coordinate-based networks have emerged as a powerful tool for 3D representation and scene reconstruction. These networks are trained to map continuous input coordinates to the value of a signal at each point. Still, current architectures are black boxes: their spectral characteristics cannot be easily analyzed, and their behavior at unsupervised points is difficult to predict. Moreover, these networks are typically trained to represent a signal at a single scale, so naive downsampling or upsampling results in artifacts. We introduce band-limited coordinate networks (BACON), a network architecture with an analytical Fourier spectrum. BACON has constrained behavior at unsupervised points, can be designed based on the spectral characteristics of the represented signal, and can represent signals at multiple scales without per-scale supervision. We demonstrate BACON for multiscale neural representation of images, radiance fields, and 3D scenes using signed distance functions and show that it outperforms conventional single-scale

coordinate networks in terms of interpretability and quality.
****
Regional Semantic Contrast and Aggregation for Weakly Supervised Semantic Segmentation
Tianfei Zhou, Meijie Zhang, Fang Zhao, Jianwu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4299-4309
Learning semantic segmentation from weakly-labeled (e.g., image tags only) data is challenging since it is hard to infer dense object regions from sparse semantic tags. Despite being broadly studied, most current efforts directly learn from limited semantic annotations carried by individual image or image pairs, and struggle to obtain integral localization maps. Our work alleviates this from a novel perspective, by exploring rich semantic contexts synergistically among abundant weakly-labeled training data for network learning and inference. In particular, we propose regional semantic contrast and aggregation (RCA). RCA is equipped with a regional memory bank to store massive, diverse object patterns appearing in training data, which acts as strong support for exploration of dataset-level semantic structure. Particularly, we propose i) semantic contrast to drive network learning by contrasting massive categorical object regions, leading to a more holistic object pattern understanding, and ii) semantic aggregation to gather diverse relational contexts in the memory to enrich semantic representations. In this manner, RCA earns a strong capability of fine-grained semantic understanding, and eventually establishes new state-of-the-art results on two popular benchmarks, i.e., PASCAL VOC 2012 and COCO 2014.
****
Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation
Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 969-978
Extracting class activation maps (CAM) is arguably the most standard step of generating pseudo masks for weakly-supervised semantic segmentation (WSSS). Yet, we find that the crux of the unsatisfactory pseudo masks is the binary cross-entropy loss (BCE) widely used in CAM. Specifically, due to the sum-over-class pooling nature of BCE, each pixel in CAM may be responsive to multiple classes co-occurring in the same receptive field. As a result, given a class, its hot CAM pixels may wrongly invade the area belonging to other classes, or the non-hot ones may be actually a part of the class. To this end, we introduce an embarrassingly simple yet surprisingly effective method: Reactivating the converged CAM with BCE by using softmax cross-entropy loss (SCE), dubbed ReCAM. Given an image, we use CAM to extract the feature pixels of every single class, and use them with the class label to learn another fully-connected layer (after the backbone) with SCE. Once converged, we extract ReCAM in the same way as in CAM. Thanks to the contrastive nature of SCE, the pixel response is disentangled into different classes and hence less mask ambiguity is expected. The evaluation on both PASCAL VOC and MS COCO shows that ReCAM can not only generates high-quality masks, but also supports plug-and-play in any CAM variant with little overhead.
****
TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions
Jeya Maria Jose Valanarasu, Rajeev Yasarla, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2353-2363
Removing adverse weather conditions like rain, fog, and snow from images is an important problem in many applications. Most methods proposed in the literature have been designed to deal with just removing one type of degradation. Recently, a CNN-based method using neural architecture search (All-in-One) was proposed to remove all the weather conditions at once. However, it has a large number of parameters as it uses multiple encoders to cater to each weather removal task and still has scope for improvement in its performance. In this work, we focus on developing an efficient solution for the all adverse weather removal problem. To this end, we propose TransWeather, a transformer-based end-to-end model with just

a single encoder and a decoder that can restore an image degraded by any weathe
r condition. Specifically, we utilize a novel transformer encoder using intra-pa
tch transformer blocks to enhance attention inside the patches to effectively re
move smaller weather degradations. We also introduce a transformer decoder with
learnable weather type embeddings to adjust to the weather degradation at hand.
TransWeather achieves significant improvements across multiple test datasets ove
r both All-in-One network as well as methods fine-tuned for specific tasks. Tran
sWeather is also validated on real world test images and found to be more effect
ive than previous methods. Implementation code can be found in the supplementary
document. Code is available at https://github.com/jeya-maria-jose/TransWeather.
****

Merry Go Round: Rotate a Frame and Fool a DNN
Daksh Thapar, Aditya Nigam, Chetan Arora; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15054-15063
A large proportion of videos captured today are first per-son videos shot from w
earable cameras. Similar to other computer vision tasks, Deep Neural Networks (D
NNs) are the workhorse for most state-of-the-art (SOTA) egocentric vision techni
ques. On the other hand DNNs are known to be susceptible to Adversarial Attacks
(AAs) which add im-perceptible noise to the input. Both black-box, as well as wh
ite-box attacks on image as well as video analysis tasks have been shown. We obs
erve that most AA techniques basically add intensity perturbation to an image. E
ven for videos, the same process is essentially repeated for each frame independ
ently. We note that the definition of imperceptibility used for images may not b
e applicable for videos, where a small intensity change happening randomly in tw
o consecutive frames may still be perceptible. In this paper we make a key novel
suggestion to use perturbation in optical flow to carry out AAs on a video anal
ysis system. Such perturbation is especially useful for egocentric videos, becau
se there is a lot of shake in the egocentric videos anyways, and adding a little
more, keeps it highly imperceptible. In general, our idea can be seen as adding
structured, para-metric noise as the adversarial perturbation. Our implementati
on of the idea by adding 3D rotations to the frames reveal that using our techni
que, one can mount a black-box AA on an egocentric activity detection system in
one-third of the queries compared to the SOTA AA technique.
****

H2FA R-CNN: Holistic and Hierarchical Feature Alignment for Cross-Domain Weakly
Supervised Object Detection
Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, Yi Yang; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 143
29-14339
Cross-domain weakly supervised object detection (CDWSOD) aims to adapt the detec
tion model to a novel target domain with easily acquired image-level annotations
. How to align the source and target domains is critical to the CDWSOD accuracy.
Existing methods usually focus on partial detection components for domain align
ment. In contrast, this paper considers that all the detection components are im
portant and proposes a Holistic and Hierarchical Feature Alignment (H^2FA) R-CNN
. H^2FA R-CNN enforces two image-level alignments for the backbone features, as
well as two instance-level alignments for the RPN and detection head. This coars
e-to-fine aligning hierarchy is in pace with the detection pipeline, i.e., proce
ssing the image-level feature and the instance-level features from bottom to top
. Importantly, we devise a novel hybrid supervision method for learning two inst
ance-level alignments. It enables the RPN and detection head to simultaneously r
eceive weak/full supervision from the target/source domains. Combining all these
feature alignments, H^2FA R-CNN effectively mitigates the gap between the sourc
e and target domains. Experimental results show that H^2FA R-CNN significantly i
mproves cross-domain object detection accuracy and sets new state of the art on
popular benchmarks. Code and pre-trained models are available at https://github.
com/XuYunqiu/H2FA_R-CNN.
****

Modeling sRGB Camera Noise With Normalizing Flows
Shayan Kousha, Ali Maleky, Michael S. Brown, Marcus A. Brubaker; Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17463-17471
Noise modeling and reduction are fundamental tasks in low-level computer vision. They are particularly important for smartphone cameras relying on small sensors that exhibit visually noticeable noise. There has recently been renewed interest in using data-driven approaches to improve camera noise models via neural networks. These data-driven approaches target noise present in the raw-sensor image before it has been processed by the camera's image signal processor (ISP). Modeling noise in the RAW-rgb domain is useful for improving and testing the in-camera denoising algorithm; however, there are situations where the camera's ISP does not apply denoising or additional denoising is desired when the RAW-rgb domain image is no longer available. In such cases, the sensor noise propagates through the ISP to the final rendered image encoded in standard RGB (sRGB). The nonlinear steps on the ISP culminate in a significantly more complex noise distribution in the sRGB domain and existing raw-domain noise models are unable to capture the sRGB noise distribution. We propose a new sRGB-domain noise model based on normalizing flows that is capable of learning the complex noise distribution found in sRGB images under various ISO levels. Our normalizing flows-based approach outperforms other models by a large margin in noise modeling and synthesis tasks. We also show that image denoisers trained on noisy images synthesized with our noise model outperforms those trained with noise from baselines models.
****

A ConvNet for the 2020s
Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976-11986
The "Roaring 20s" of visual recognition began with the introduction of Vision Transformers (ViTs), which quickly superseded ConvNets as the state-of-the-art image classification model. A vanilla ViT, on the other hand, faces difficulties when applied to general computer vision tasks such as object detection and semantic segmentation. It is the hierarchical Transformers (e.g., Swin Transformers) that reintroduced several ConvNet priors, making Transformers practically viable as a generic vision backbone and demonstrating remarkable performance on a wide variety of vision tasks. However, the effectiveness of such hybrid approaches is still largely credited to the intrinsic superiority of Transformers, rather than the inherent inductive biases of convolutions. In this work, we reexamine the design spaces and test the limits of what a pure ConvNet can achieve. We gradually "modernize" a standard ResNet toward the design of a vision Transformer, and discover several key components that contribute to the performance difference along the way. The outcome of this exploration is a family of pure ConvNet models dubbed ConvNeXt. Constructed entirely from standard ConvNet modules, ConvNeXts compete favorably with Transformers in terms of accuracy and scalability, achieving 87.8% ImageNet top-1 accuracy and outperforming Swin Transformers on COCO detection and ADE20K segmentation, while maintaining the simplicity and efficiency of standard ConvNets.
****

Reference-Based Video Super-Resolution Using Multi-Camera Video Triplets
Junyong Lee, Myeonghee Lee, Sunghyun Cho, Seungyong Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17824-17833
We propose the first reference-based video super-resolution (RefVSR) approach that utilizes reference videos for high-fidelity results. We focus on RefVSR in a triple-camera setting, where we aim at super-resolving a low-resolution ultra-wide video utilizing wide-angle and telephoto videos. We introduce the first RefVSR network that recurrently aligns and propagates temporal reference features fused with features extracted from low-resolution frames. To facilitate the fusion and propagation of temporal reference features, we propose a propagative temporal fusion module. For learning and evaluation of our network, we present the first RefVSR dataset consisting of triplets of ultra-wide, wide-angle, and telephoto videos concurrently taken from triple cameras of a smartphone. We also propose

a two-stage training strategy fully utilizing video triplets in the proposed dat
aset for real-world 4x video super-resolution. We extensively evaluate our metho
d, and the result shows the state-of-the-art performance in 4x super-resolution.
****

Self-Supervised Image Representation Learning With Geometric Set Consistency
Nenglun Chen, Lei Chu, Hao Pan, Yan Lu, Wenping Wang; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19292-
19302
We propose a method for self-supervised image representation learning under the
guidance of 3D geometric consistency. Our intuition is that 3D geometric consist
ency priors such as smooth regions and surface discontinuities may imply consist
ent semantics or object boundaries, and can act as strong cues to guide the lear
ning of 2D image representations without semantic labels. Specifically, we intro
duce 3D geometric consistency into a contrastive learning framework to enforce t
he feature consistency within image views. We propose to use geometric consisten
cy sets as constraints and adapt the InfoNCE loss accordingly. We show that our
learned image representations are general. By fine-tuning our pre-trained repres
entations for various 2D image-based downstream tasks, including semantic segmen
tation, object detection, and instance segmentation on real-world indoor scene d
atasets, we achieve superior performance compared with state-of-the-art methods.
****

Deep Anomaly Discovery From Unlabeled Videos via Normality Advantage and Self-Pa
ced Refinement
Guang Yu, Siqi Wang, Zhiping Cai, Xinwang Liu, Chuanfu Xu, Chengkun Wu; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2022, pp. 13987-13998
While classic video anomaly detection (VAD) requires labeled normal videos for t
raining, emerging unsupervised VAD (UVAD) aims to discover anomalies directly fr
om fully unlabeled videos. However, existing UVAD methods still rely on shallow
models to perform detection or initialization, and they are evidently inferior t
o classic VAD methods. This paper proposes a full deep neural network (DNN) base
d solution that can realize highly effective UVAD. First, we, for the first time
, point out that deep reconstruction can be surprisingly effective for UVAD, whi
ch inspires us to unveil a property named "normality advantage", i.e., normal ev
ents will enjoy lower reconstruction loss when DNN learns to reconstruct unlabel
ed videos. With this property, we propose Localization based Reconstruction (LBR
) as a strong UVAD baseline and a solid foundation of our solution. Second, we p
ropose a novel self-paced refinement (SPR) scheme, which is synthesized into LBR
 to conduct UVAD. Unlike ordinary self-paced learning that injects more samples
in an easy-to-hard manner, the proposed SPR scheme gradually drops samples so th
at suspicious anomalies can be removed from the learning process. In this way, S
PR consolidates normality advantage and enables better UVAD in a more proactive
way. Finally, we further design a variant solution that explicitly takes the mot
ion cues into account. The solution evidently enhances the UVAD performance, and
 it sometimes even surpasses the best classic VAD methods. Experiments show that
 our solution not only significantly outperforms existing UVAD methods by a wide
 margin (5% to 9% AUROC), but also enables UVAD to catch up with the mainstream
performance of classic VAD.
****

P3Depth: Monocular Depth Estimation With a Piecewise Planarity Prior
Vaishakh Patil, Christos Sakaridis, Alexander Liniger, Luc Van Gool; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
022, pp. 1610-1621
Monocular depth estimation is vital for scene understanding and downstream tasks
. We focus on the supervised setup, in which ground-truth depth is available onl
y at training time. Based on knowledge about the high regularity of real 3D scen
es, we propose a method that learns to selectively leverage information from cop
lanar pixels to improve the predicted depth. In particular, we introduce a piece
wise planarity prior which states that for each pixel, there is a seed pixel whi
ch shares the same planar 3D surface with the former. Motivated by this prior, w

e design a network with two heads. The first head outputs pixel-level plane coef
ficients, while the second one outputs a dense offset vector field that identifi
es the positions of seed pixels. The plane coefficients of seed pixels are then
used to predict depth at each position. The resulting prediction is adaptively f
used with the initial prediction from the first head via a learned confidence to
account for potential deviations from precise local planarity. The entire archi
tecture is trained end-to-end thanks to the differentiability of the proposed mo
dules and it learns to predict regular depth maps, with sharp edges at occlusion
boundaries. An extensive evaluation of our method shows that we set the new sta
te of the art in supervised monocular depth estimation, surpassing prior methods
on NYU Depth-v2 and on the Garg split of KITTI. Our method delivers depth maps
that yield plausible 3D reconstructions of the input scenes. Code is available a
t: https://github.com/SysCV/P3Depth
****

GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Det
ection
Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, Si Liu; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 20123-20132
The task of Human-Object Interaction (HOI) detection could be divided into two c
ore problems, i.e., human-object association and interaction understanding. In t
his paper, we reveal and address the disadvantages of the conventional query-dri
ven HOI detectors from the two aspects. For the association, previous two-branch
methods suffer from complex and costly post-matching, while single-branch metho
ds ignore the features distinction in different tasks. We propose Guided-Embeddi
ng Network (GEN) to attain a two-branch pipeline without post-matching. In GEN,
we design an instance decoder to detect humans and objects with two independent
query sets and a position Guided Embedding (p-GE) to mark the human and object i
n the same position as a pair. Besides, we design an interaction decoder to clas
sify interactions, where the interaction queries are made of instance Guided Emb
eddings (i-GE) generated from the outputs of each instance decoder layer. For th
e interaction understanding, previous methods suffer from long-tailed distributi
on and zero-shot discovery. This paper proposes a Visual-Linguistic Knowledge Tr
ansfer (VLKT) training strategy to enhance interaction understanding by transfer
ring knowledge from a visual-linguistic pre-trained model CLIP. In specific, we
extract text embeddings for all labels with CLIP to initialize the classifier an
d adopt a mimic loss to minimize the visual feature distance between GEN and CLI
P. As a result, GEN-VLKT outperforms the state of the art by large margins on mu
ltiple datasets, e.g., +5.05 mAP on HICO-Det. The source codes are available at
https://github.com/YueLiao/gen-vlkt.
****

Simple Multi-Dataset Detection
Xingyi Zhou, Vladlen Koltun, Philipp Krähenbühl; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7571-7580
How do we build a general and broad object detection system? We use all labels o
f all concepts ever annotated. These labels span diverse datasets with potential
ly inconsistent taxonomies. In this paper, we present a simple method for traini
ng a unified detector on multiple large-scale datasets. We use dataset-specific
training protocols and losses, but share a common detection architecture with da
taset-specific outputs. We show how to automatically integrate these dataset-spe
cific outputs into a common semantic taxonomy. In contrast to prior work, our ap
proach does not require manual taxonomy reconciliation. Experiments show our lea
rned taxonomy outperforms a expert-designed taxonomy in all datasets. Our multi-
dataset detector performs as well as dataset-specific models on each training do
main, and can generalize to new unseen dataset without fine-tuning on them. Code
is available at https://github.com/xingyizhou/UniDet.
****

MLP-3D: A MLP-Like 3D Architecture With Grouped Time Mixing
Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Tao Mei; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3062-3072

Convolutional Neural Networks (CNNs) have been regarded as the go-to models for visual recognition. More recently, convolution-free networks, based on multi-head self-attention (MSA) or multi-layer perceptrons (MLPs), become more and more popular. Nevertheless, it is not trivial when utilizing these newly-minted networks for video recognition due to the large variations and complexities in video data. In this paper, we present MLP-3D networks, a novel MLP-like 3D architecture for video recognition. Specifically, the architecture consists of MLP-3D blocks, where each block contains one MLP applied across tokens (i.e., token-mixing MLP) and one MLP applied independently to each token (i.e., channel MLP). By deriving the novel grouped time mixing (GTM) operations, we equip the basic token-mixing MLP with the ability of temporal modeling. GTM divides the input tokens into several temporal groups and linearly maps the tokens in each group with the shared projection matrix. Furthermore, we devise several variants of GTM with different grouping strategies, and compose each variant in different blocks of MLP-3D network by greedy architecture search. Without the dependence on convolutions or attention mechanisms, our MLP-3D networks achieves 68.5%/81.4% top-1 accuracy on Something-Something V2 and Kinetics-400 datasets, respectively. Despite with fewer computations, the results are comparable to state-of-the-art widely-used 3D CNNs and video transformers.

****

Proactive Image Manipulation Detection

Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15386-15395

Image manipulation detection algorithms are often trained to discriminate between images manipulated with particular Generative Models (GMs) and genuine/real images, yet generalize poorly to images manipulated with GMs unseen in the training. Conventional detection algorithms receive an input image passively. By contrast, we propose a proactive scheme to image manipulation detection. Our key enabling technique is to estimate a set of templates which when added onto the real image would lead to more accurate manipulation detection. That is, a template protected real image, and its manipulated version, is better discriminated compared to the original real image vs. its manipulated one. These templates are estimated using certain constraints based on the desired properties of templates. For image manipulation detection, our proposed approach outperforms the prior work by an average precision of 16% for CycleGAN and 32% for GauGAN. Our approach is generalizable to a variety of GMs showing an improvement over prior work by an average precision of 10% averaged across 12 GMs. Our code is available at https://www.github.com/vishal3477/proactive_IMD.

****

Sketch3T: Test-Time Training for Zero-Shot SBIR

Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7462-7471

Zero-shot sketch-based image retrieval typically asks for a trained model to be applied as is to unseen categories. In this paper, we question to argue that this setup by definition is not compatible with the inherent abstract and subjective nature of sketches -- the model might transfer well to new categories, but will not understand sketches existing in different test-time distribution as a result. We thus extend ZS-SBIR asking it to transfer to both categories and sketch distributions. Our key contribution is a test-time training paradigm that can adapt using just one sketch. Since there is no paired photo, we make use of a sketch raster-vector reconstruction module as a self-supervised auxiliary task. To maintain the fidelity of the trained cross-modal joint embedding during test-time update, we design a novel meta-learning based training paradigm to learn a separation between model updates incurred by this auxiliary task from those off the primary objective of discriminative learning. Extensive experiments show our model to outperform state-of-the-arts, thanks to the proposed test-time adaption that not only transfers to new categories but also accommodates to new sketching styles.

****
BANMo: Building Animatable 3D Neural Models From Many Casual Videos
Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, Hanbyul Joo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2863-2873
Prior work for articulated 3D shape reconstruction often relies on specialized multi-view and depth sensors or pre-built deformable 3D models. Such methods do not scale to diverse sets of objects in the wild. We present a method that requires neither of them. It builds high-fidelity, articulated 3D models from many monocular casual videos in a differentiable rendering framework. Our key insight is to merge three schools of thought: (1) classic deformable shape models that make use of articulated bones and blend skinning, (2) canonical embeddings that establish correspondences between pixels and a canonical 3D model, and (3) volumetric neural radiance fields (NeRFs) that are amenable to gradient-based optimization. We introduce neural blend skinning models that allow for differentiable and invertible articulated deformations. When combined with canonical embeddings, such models allow us to establish dense correspondences across videos that can be self-supervised with cycle consistency. On real and synthetic datasets, our method shows higher-fidelity 3D reconstructions than prior works for humans and animals, with the ability to render realistic images from novel viewpoints.
****
StyTr2: Image Style Transfer With Transformers
Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11326-11336
The goal of image style transfer is to render an image with artistic features guided by a style reference while maintaining the original content. Owing to the locality in convolutional neural networks (CNNs), extracting and maintaining the global information of input images is difficult. Therefore, traditional neural style transfer methods face biased content representation. To address this critical issue, we take long-range dependencies of input images into account for image style transfer by proposing a transformer-based approach called StyTr^2. In contrast with visual transformers for other vision tasks, StyTr^2 contains two different transformer encoders to generate domain-specific sequences for content and style, respectively. Following the encoders, a multi-layer transformer decoder is adopted to stylize the content sequence according to the style sequence. We also analyze the deficiency of existing positional encoding methods and propose the content-aware positional encoding (CAPE), which is scale-invariant and more suitable for image style transfer tasks. Qualitative and quantitative experiments demonstrate the effectiveness of the proposed StyTr^2 compared with state-of-the-art CNN-based and flow-based approaches. Code and models are available at https://github.com/diyiiyiii/StyTR-2.
****
Towards Discriminative Representation: Multi-View Trajectory Contrastive Learning for Online Multi-Object Tracking
En Yu, Zhuoling Li, Shoudong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8834-8843
Discriminative representation is crucial for the association step in multi-object tracking. Recent work mainly utilizes features in single or neighboring frames for constructing metric loss and empowering networks to extract representation of targets. Although this strategy is effective, it fails to fully exploit the information contained in a whole trajectory. To this end, we propose a strategy, namely multi-view trajectory contrastive learning, in which each trajectory is represented as a center vector. By maintaining all the vectors in a dynamically updated memory bank, a trajectory-level contrastive loss is devised to explore the inter-frame information in the whole trajectories. Besides, in this strategy, each target is represented as multiple adaptively selected keypoints rather than a pre-defined anchor or center. This design allows the network to generate richer representation from multiple views of the same target, which can better characterize occluded objects. Additionally, in the inference stage, a similarity-gui

ded feature fusion strategy is developed for further boosting the quality of the trajectory representation. Extensive experiments have been conducted on MOTChallenge to verify the effectiveness of the proposed techniques. The experimental results indicate that our method has surpassed preceding trackers and established new state-of-the-art performance.

****

Global Matching With Overlapping Attention for Optical Flow Estimation

Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, Dimitris Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17592-17601

Optical flow estimation is a fundamental task in computer vision. Recent direct-regression methods using deep neural networks achieve remarkable performance improvement. However, they do not explicitly capture long-term motion correspondences and thus cannot handle large motions effectively. In this paper, inspired by the traditional matching-optimization methods where matching is introduced to handle large displacements before energy-based optimizations, we introduce a simple but effective global matching step before the direct regression and develop a learning-based matching-optimization framework, namely GMFlowNet. In GMFlowNet, global matching is efficiently calculated by applying argmax on 4D cost volumes. Additionally, to improve the matching quality, we propose patch-based overlapping attention to extract large context features. Extensive experiments demonstrate that GMFlowNet outperforms RAFT, the most popular optimization-only method, by a large margin and achieves state-of-the-art performance on standard benchmarks. Thanks to the matching and overlapping attention, GMFlowNet obtains major improvements on the predictions for textureless regions and large motions. Our code is made publicly available at https://github.com/xiaofeng94/GMFlowNet.

****

Language As Queries for Referring Video Object Segmentation

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4974-4984

Referring video object segmentation (R-VOS) is an emerging cross-modal task that aims to segment the target object referred by a language expression in all video frames. In this work, we propose a simple and unified framework built upon Transformer, termed ReferFormer. It views the language as queries and directly attends to the most relevant regions in the video frames. Concretely, we introduce a small set of object queries conditioned on the language as the input to the Transformer. In this manner, all the queries are obligated to find the referred objects only. They are eventually transformed into dynamic kernels which capture the crucial object-level information, and play the role of convolution filters to generate the segmentation masks from feature maps. The object tracking is achieved naturally by linking the corresponding queries across frames. This mechanism greatly simplifies the pipeline and the endto-end framework is significantly different from the previous methods. Extensive experiments on Ref-Youtube-VOS, Ref-DAVIS17, A2D-Sentences and JHMDB-Sentences show the effectiveness of ReferFormer. On Ref-Youtube-VOS, ReferFormer achieves 55.6 J &F with a ResNet-50 backbone without bells and whistles, which exceeds the previous state-of-the-art performance by 8.4 points. In addition, with the strong Video-Swin-Base backbone, ReferFormer achieves the best J &F of 64.9 among all existing methods. Moreover, we show the impressive results of 55.0 mAP and 43.7 mAP on A2D-Sentences and JHMDB-Sentences respectively, which significantly outperforms the previous methods by a large margin. Code is publicly available at https://github.com/wjn922/ReferFormer.

****

Investigating the Impact of Multi-LiDAR Placement on Object Detection for Autonomous Driving

Hanjiang Hu, Zuxin Liu, Sharad Chitlangia, Akhil Agnihotri, Ding Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2550-2559

The past few years have witnessed an increasing interest in improving the percep

tion performance of LiDARs on autonomous vehicles. While most of the existing wo rks focus on developing new deep learning algorithms or model architectures, we study the problem from the physical design perspective, i.e., how different plac ements of multiple LiDARs influence the learning-based perception. To this end, we introduce an easy-to-compute information-theoretic surrogate metric to quanti tatively and fast evaluate LiDAR placement for 3D detection of different types o f objects. We also present a new data collection, detection model training and e valuation framework in the realistic CARLA simulator to evaluate disparate multi -LiDAR configurations. Using several prevalent placements inspired by the design s of self-driving companies, we show the correlation between our surrogate metri c and object detection performance of different representative algorithms on KIT TI through extensive experiments, validating the effectiveness of our LiDAR plac ement evaluation approach. Our results show that sensor placement is non-negligi ble in 3D point cloud-based object detection, which will contribute to 5% 10% performance discrepancy in terms of average precision in challenging 3D object d etection settings. We believe that this is one of the first studies to quantitat ively investigate the influence of LiDAR placement on perception performance.
****

MViTv2: Improved Multiscale Vision Transformers for Classification and Detection
Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Mal ik, Christoph Feichtenhofer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4804-4814
In this paper, we study Multiscale Vision Transformers (MViTv2) as a unified arc hitecture for image and video classification, as well as object detection. We pr esent an improved version of MViT that incorporates decomposed relative position al embeddings and residual pooling connections. We instantiate this architecture in five sizes and evaluate it for ImageNet classification, COCO detection and K inetics video recognition where it outperforms prior work. We further compare MV iTv2s' pooling attention to window attention mechanisms where it outperforms the latter in accuracy/compute. Without bells-and-whistles, MViTv2 has state-of-the -art performance in 3 domains: 88.8% accuracy on ImageNet classification, 58.7 b oxAP on COCO object detection as well as 86.1% on Kinetics-400 video classificat ion. Code and models are available at https://github.com/facebookresearch/mvit.
****

Audio-Visual Generalised Zero-Shot Learning With Cross-Modal Attention and Langu age
Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 10553-10563
Learning to classify video data from classes not included in the training data, i.e. video-based zero-shot learning, is challenging. We conjecture that the natu ral alignment between the audio and visual modalities in video data provides a r ich training signal for learning discriminative multi-modal representations. Foc using on the relatively underexplored task of audio-visual zero-shot learning, w e propose to learn multi-modal representations from audio-visual data using cros s-modal attention and exploit textual label embeddings for transferring knowledg e from seen classes to unseen classes. Taking this one step further, in our gene ralised audio-visual zero-shot learning setting, we include all the training cla sses in the test-time search space which act as distractors and increase the dif ficulty while making the setting more realistic. Due to the lack of a unified be nchmark in this domain, we introduce a (generalised) zero-shot learning benchmar k on three audio-visual datasets of varying sizes and difficulty, VGGSound, UCF, and ActivityNet, ensuring that the unseen test classes do not appear in the dat aset used for supervised training of the backbone deep models. Comparing multipl e relevant and recent methods, we demonstrate that our proposed AVCA model achie ves state-of-the-art performance on all three datasets. Code and data are availa ble at https://github.com/ExplainableML/AVCA-GZSL.
****

Rethinking Efficient Lane Detection via Curve Modeling
Zhengyang Feng, Shaohua Guo, Xin Tan, Ke Xu, Min Wang, Lizhuang Ma; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 17062-17070
This paper presents a novel parametric curve-based method for lane detection in
RGB images. Unlike state-of-the-art segmentation-based and point detection-based
 methods that typically require heuristics to either decode predictions or formu
late a large sum of anchors, the curve-based methods can learn holistic lane rep
resentations naturally. To handle the optimization difficulties of existing poly
nomial curve methods, we propose to exploit the parametric Bezier curve due to i
ts ease of computation, stability, and high freedom degrees of transformations.
In addition, we propose the deformable convolution-based feature flip fusion, fo
r exploiting the symmetry properties of lanes in driving scenes. The proposed me
thod achieves a new state-of-the-art performance on the popular LLAMAS benchmark
. It also achieves favorable accuracy on the TuSimple and CULane datasets, while
 retaining both low latency (> 150 FPS) and small model size (< 10M). Our method
 can serve as a new baseline, to shed the light on the parametric curves modelin
g for lane detection. Codes of our model and PytorchAutoDrive: a unified framewo
rk for self-driving perception, are available at: https://github.com/voldemortX/
pytorch-auto-drive.
****

GreedyNASv2: Greedier Search With a Greedy Path Filter
Tao Huang, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, Chang
Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2022, pp. 11902-11911
Training a good supernet in one-shot NAS methods is difficult since the search s
pace is usually considerably huge (e.g., 13^ 21). In order to enhance the super
net's evaluation ability, one greedy strategy is to sample good paths, and let t
he supernet lean towards the good ones and ease its evaluation burden as a resul
t. However, in practice the search can be still quite inefficient since the iden
tification of good paths is not accurate enough and sampled paths still scatter
around the whole search space. In this paper, we leverage an explicit path filte
r to capture the characteristics of paths and directly filter those weak ones, s
o that the search can be thus implemented on the shrunk space more greedily and
efficiently. Concretely, based on the fact that good paths are way much less tha
n the weak ones in the space, we argue that the label of "weak paths" will be mo
re confident and reliable than that of "good paths" in multi-path sampling. In t
his way, we thus cast the training of path filter in the positive and unlabeled
(PU) learning paradigm, and also encourage a path embedding as better path/opera
tion representation to enhance the identification capacity of the learned filter
. By dint of this embedding, we can further shrink the search space by aggregati
ng similar operations with similar embeddings, and the search can be more effici
ent and accurate. Extensive experiments validate the effectiveness of the propos
ed method GreedyNASv2. For example, our obtained GreedyNASv2-L achieves 81.1% To
p-1 accuracy on ImageNet dataset, significantly outperforming the ResNet-50 stro
ng baselines.
****

Self-Supervised Arbitrary-Scale Point Clouds Upsampling via Implicit Neural Repr
esentation
Wenbo Zhao, Xianming Liu, Zhiwei Zhong, Junjun Jiang, Wei Gao, Ge Li, Xiangyang
Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2022, pp. 1999-2007
Point clouds upsampling is a challenging issue to generate dense and uniform poi
nt clouds from the given sparse input. Most existing methods either take the end
-to-end supervised learning based manner, where large amounts of pairs of sparse
 input and dense ground-truth are exploited as supervision information; or treat
 up-scaling of different scale factors as independent tasks, and have to build m
ultiple networks to handle upsampling with varying factors. In this paper, we pr
opose a novel approach that achieves selfsupervised and magnification-flexible p
oint clouds upsampling simultaneously. We formulate point clouds upsampling as t
he task of seeking nearest projection points on the implicit surface for seed po
ints. To this end, we define two implicit neural functions to estimate projectio

n direction and distance respectively, which can be trained by two pretext learn ing tasks. Experimental results demonstrate that our self-supervised learning ba sed scheme achieves competitive or even better performance than supervised learn ing based state-of-the-art methods. The source code is publicly available at htt ps://github.com/xnowbzhao/sapcu.

****

Co-Advise: Cross Inductive Bias Distillation

Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, Ha ng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 16773-16782

The inductive bias of vision transformers is more relaxed that cannot work well with insufficient data. Knowledge distillation is thus introduced to assist the training of transformers. Unlike previous works, where merely heavy convolution-based teachers are provided, in this paper, we delve into the influence of model s inductive biases in knowledge distillation (e.g., convolution and involution). Our key observation is that the teacher accuracy is not the dominant reason for the student accuracy, but the teacher inductive bias is more important. We demo nstrate that lightweight teachers with different architectural inductive biases can be used to co-advise the student transformer with outstanding performances. The rationale behind is that models designed with different inductive biases ten d to focus on diverse patterns, and teachers with different inductive biases att ain various knowledge despite being trained on the same dataset. The diverse kno wledge provides a more precise and comprehensive description of the data and com pounds and boosts the performance of the student during distillation. Furthermor e, we propose a token inductive bias alignment to align the inductive bias of th e token with its target teacher model. With only lightweight teachers provided a nd using this cross inductive bias distillation method, our vision transformers (termed as CiT) outperform all previous vision transformers (ViT) of the same ar chitecture on ImageNet. Moreover, our small size model CiT-SAK further achieves 82.7% Top-1 accuracy on ImageNet without modifying the attention module of the V iT. Code is available at https://github.com/OliverRensu/co-advise

****

AdaMixer: A Fast-Converging Query-Based Object Detector

Ziteng Gao, Limin Wang, Bing Han, Sheng Guo; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5364-5373

Traditional object detectors employ the dense paradigm of scanning over location s and scales in an image. The recent query-based object detectors break this con vention by decoding image features with a set of learnable queries. However, thi s paradigm still suffers from slow convergence, limited performance, and design complexity of extra networks between backbone and decoder. In this paper, we fin d that the key to these issues is the adaptability of decoders for casting queri es to varying objects. Accordingly, we propose a fast-converging query-based det ector, named AdaMixer, by improving the adaptability of query-based decoding pro cesses in two aspects. First, each query adaptively samples features over space and scales based on estimated offsets, which allows AdaMixer to efficiently atte nd to the coherent regions of objects. Then, we dynamically decode these sampled features with an adaptive MLP-Mixer under the guidance of each query. Thanks to these two critical designs, AdaMixer enjoys architectural simplicity without re quiring dense attentional encoders or explicit pyramid networks. On the challeng ing MS COCO benchmark, AdaMixer with ResNet-50 as the backbone, with 12 training epochs, reaches up to 45.0 AP on the validation set along with 27.9 APs in dete cting small objects. With the longer training scheme, AdaMixer with ResNeXt-101-DCN and Swin-S reaches 49.5 and 51.3 AP. Our work sheds light on a simple, accur ate, and fast converging architecture for query-based object detectors. The code is made available at https://github.com/MCG-NJU/AdaMixer.

****

DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histop athology Whole Slide Image Classification

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coup land, Yalin Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), 2022, pp. 18802-18812
Multiple instance learning (MIL) has been increasingly used in the classification of histopathology whole slide images (WSIs). However, MIL approaches for this specific classification problem still face unique challenges, particularly those related to small sample cohorts. In these, there are limited number of WSI slides (bags), while the resolution of a single WSI is huge, which leads to a large number of patches (instances) cropped from this slide. To address this issue, we propose to virtually enlarge the number of bags by introducing the concept of pseudo-bags, on which a double-tier MIL framework is built to effectively use the intrinsic features. Besides, we also contribute to deriving the instance probability under the framework of attention-based MIL, and utilize the derivation to help construct and analyze the proposed framework. The proposed method outperforms other latest methods on the CAMELYON-16 by substantially large margins, and is also better in performance on the TCGA lung cancer dataset. The proposed framework is ready to be extended for wider MIL applications. The code is available at: https://github.com/hrzhang1123/DTFD-MIL
****

BEVT: BERT Pretraining of Video Transformers
Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, Lu Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14733-14743
This paper studies the BERT pretraining of video transformers. It is a straightforward but worth-studying extension given the recent success from BERT pretraining of image transformers. We introduce BEVT which decouples video representation learning into spatial representation learning and temporal dynamics learning. In particular, BEVT first performs masked image modeling on image data, and then conducts masked image modeling jointly with masked video modeling on video data. This design is motivated by two observations: 1) transformers learned on image datasets provide decent spatial priors that can ease the learning of video transformers, which are often times computationally-intensive if trained from scratch; 2) discriminative clues, i.e., spatial and temporal information, needed to make correct predictions vary among different videos due to large intra-class and inter-class variations. We conduct extensive experiments on three challenging video benchmarks where BEVT achieves very promising results. On Kinetics 400, for which recognition mostly relies on discriminative spatial representations, BEVT achieves comparable results to strong supervised baselines. On Something-Something-V2 and Diving 48, which contain videos relying on temporal dynamics, BEVT outperforms by clear margins all alternative baselines and achieves state-of-the-art performance with a 71.4% and 87.2% Top-1 accuracy respectively. Code is available at https://github.com/xyzforever/BEVT.
****

Deep Generalized Unfolding Networks for Image Restoration
Chong Mou, Qian Wang, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17399-17410
Deep neural networks (DNN) have achieved great success in image restoration. However, most DNN methods are designed as a black box, lacking transparency and interpretability. Although some methods are proposed to combine traditional optimization algorithms with DNN, they usually demand pre-defined degradation processes or handcrafted assumptions, making it difficult to deal with complex and real-world applications. In this paper, we propose a Deep Generalized Unfolding Network (DGUNet) for image restoration. Concretely, without loss of interpretability, we integrate a gradient estimation strategy into the gradient descent step of the Proximal Gradient Descent (PGD) algorithm, driving it to deal with complex and real-world image degradation. In addition, we design inter-stage information pathways across proximal mapping in different PGD iterations to rectify the intrinsic information loss in most deep unfolding networks (DUN) through a multi-scale and spatial-adaptive way. By integrating the flexible gradient descent and informative proximal mapping, we unfold the iterative PGD algorithm into a trainable DNN. Extensive experiments on various image restoration tasks demonstrate the superiority of our method in terms of state-of-the-art performance, interpretabil

ity, and generalizability. The source code is available at https://github.com/MC-E/Deep-Generalized-Unfolding-Networks-for-Image-Restoration.
****

Automatic Relation-Aware Graph Network Proliferation
Shaofei Cai, Liang Li, Xinzhe Han, Jiebo Luo, Zheng-Jun Zha, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10863-10873
Graph neural architecture search has sparked much attention as Graph Neural Networks (GNNs) have shown powerful reasoning capability in many relational tasks. However, the currently used graph search space overemphasizes learning node features and neglects mining hierarchical relational information. Moreover, due to diverse mechanisms in the message passing, the graph search space is much larger than that of CNNs. This hinders the straightforward application of classical search strategies for exploring complicated graph search space. We propose Automatic Relation-aware Graph Network Proliferation (ARGNP) for efficiently searching GNNs with a relation-guided message passing mechanism. Specifically, we first devise a novel dual relation-aware graph search space that comprises both node and relation learning operations. These operations can extract hierarchical node/relational information and provide anisotropic guidance for message passing on a graph. Second, analogous to cell proliferation, we design a network proliferation search paradigm to progressively determine the GNN architectures by iteratively performing network division and differentiation. The experiments on six datasets for four graph learning tasks demonstrate that GNNs produced by our method are superior to the current state-of-the-art hand-crafted and search-based GNNs.
****

AIM: An Auto-Augmenter for Images and Meshes
Vinit Veerendraveer Singh, Chandra Kambhamettu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 722-731
Data augmentations are commonly used to increase the robustness of deep neural networks. In most contemporary research, the networks do not decide the augmentations; they are task-agnostic, and grid search determines their magnitudes. Furthermore, augmentations applicable to lower-dimensional data do not easily extend to higher-dimensional data and vice versa. This paper presents an auto-augmenter for images and meshes (AIM) that easily incorporates into neural networks at training and inference times. It jointly optimizes with the network to produce constrained, non-rigid deformations in the data. AIM predicts sample-aware deformations suited for a task, and our experiments confirm its effectiveness with various networks.
****

VISOLO: Grid-Based Space-Time Aggregation for Efficient Online Video Instance Segmentation
Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2896-2905
For online video instance segmentation (VIS), fully utilizing the information from previous frames in an efficient manner is essential for real-time applications. Most previous methods follow a two-stage approach requiring additional computations such as RPN and RoIAlign, and do not fully exploit the available information in the video for all subtasks in VIS. In this paper, we propose a novel single-stage framework for online VIS built based on the grid structured feature representation. The grid-based features allow us to employ a fully convolutional networks for real-time processing, and also to easily reuse and share features within different components. We also introduce cooperatively operating modules that aggregate information from available frames, in order to enrich the features for all subtasks in VIS. Our design fully takes advantage of previous information in a grid form for all tasks in VIS in an efficient way, and we achieved the state-of-the-art accuracy (38.6 AP and 36.9 AP) and speed (40.0 FPS) on YouTube-VIS 2019 and 2021 datasets among online VIS methods. The code is available at https://github.com/SuHoHan95/VISOLO.
****

Deep Unlearning via Randomized Conditionally Independent Hessians
Ronak Mehta, Sourav Pal, Vikas Singh, Sathya N. Ravi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10422-10431
Recent legislation has led to interest in machine unlearning, i.e., removing specific training samples from a predictive model as if they never existed in the training dataset. Unlearning may also be required due to corrupted/adversarial data or simply a user's updated privacy requirement. For models which require no training (k-NN), simply deleting the closest original sample can be effective. But this idea is inapplicable to models which learn richer representations. Recent ideas leveraging optimization-based updates scale poorly with the model dimension d, due to inverting the Hessian of the loss function. We use a variant of a new conditional independence coefficient, L-CODEC, to identify a subset of the model parameters with the most semantic overlap on an individual sample level. Our approach completely avoids the need to invert a (possibly) huge matrix. By utilizing a Markov blanket selection, we premise that L-CODEC is also suitable for deep unlearning, as well as other applications in vision. Compared to alternatives, L-CODEC makes approximate unlearning possible in settings that would otherwise be infeasible, including vision models used for face recognition, person re-identification and NLP models that may require unlearning samples identified for exclusion. Code can be found at https://github.com/vsingh-group/LCODEC-deep-unlearning/
****
Patch-Level Representation Learning for Self-Supervised Vision Transformers
Sukmin Yun, Hankook Lee, Jaehyung Kim, Jinwoo Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8354-8363
Recent self-supervised learning (SSL) methods have shown impressive results in learning visual representations from unlabeled images. This paper aims to improve their performance further by utilizing the architectural advantages of the underlying neural network, as the current state-of-the-art visual pretext tasks for SSL do not enjoy the benefit, i.e., they are architecture-agnostic. In particular, we focus on Vision Transformers (ViTs), which have gained much attention recently as a better architectural choice, often outperforming convolutional networks for various visual tasks. The unique characteristic of ViT is that it takes a sequence of disjoint patches from an image and processes patch-level representations internally. Inspired by this, we design a simple yet effective visual pretext task, coined SelfPatch, for learning better patch-level representations. To be specific, we enforce invariance against each patch and its neighbors, i.e., each patch treats similar neighboring patches as positive samples. Consequently, training ViTs with SelfPatch learns more semantically meaningful relations among patches (without using human-annotated labels), which can be beneficial, in particular, to downstream tasks of a dense prediction type. Despite its simplicity, we demonstrate that it can significantly improve the performance of existing SSL methods for various visual tasks, including object detection and semantic segmentation. Specifically, SelfPatch significantly improves the recent self-supervised ViT, DINO, by achieving +1.3 AP on COCO object detection, +1.2 AP on COCO instance segmentation, and +2.9 mIoU on ADE20K semantic segmentation.
****
Sylph: A Hypernetwork Framework for Incremental Few-Shot Object Detection
Li Yin, Juan M. Perez-Rua, Kevin J. Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9035-9045
We study the challenging incremental few-shot object detection (iFSD) setting. Recently, hypernetwork-based approaches have been studied in the context of continuous and finetune-free iFSD with limited success. We take a closer look at important design choices of such methods, leading to several key improvements and resulting in a more accurate and flexible framework, which we call Sylph. In particular, we demonstrate the effectiveness of decoupling object classification from localization by leveraging a base detector that is pretrained for class-agnostic localization on large-scale dataset. Contrary to what previous results have su

ggested, we show that with a carefully designed class-conditional hypernetwork, finetune-free iFSD can be highly effective, especially when a large number of base categories with abundant data are available for meta-training, almost approaching alternatives that undergo test-time-training. This result is even more significant considering its many practical advantages: (1) incrementally learning new classes in sequence without additional training, (2) detecting both novel and seen classes in a single pass, and (3) no forgetting of previously seen classes. We benchmark our model on both COCO and LVIS, reporting as high as 17% AP on the long-tail rare classes on LVIS, indicating the promise of hypernetwork-based iFSD.

****

Incremental Learning in Semantic Segmentation From Image Labels
Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, Barbara Caputo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4371-4381
Although existing semantic segmentation approaches achieve impressive results, they still struggle to update their models incrementally as new categories are uncovered. Furthermore, pixel-by-pixel annotations are expensive and time-consuming. This paper proposes a novel framework for Weakly Incremental Learning for Semantic Segmentation, that aims at learning to segment new classes from cheap and largely available image-level labels. As opposed to existing approaches, that need to generate pseudo-labels offline, we use a localizer, trained with image-level labels and regularized by the segmentation model, to obtain pseudo-supervision online and update the model incrementally. We cope with the inherent noise in the process by using soft-labels generated by the localizer. We demonstrate the effectiveness of our approach on the Pascal VOC and COCO datasets, outperforming offline weakly-supervised methods and obtaining results comparable with incremental learning methods with full supervision.

****

Playable Environments: Video Manipulation in Space and Time
Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3584-3593
We present Playable Environments - a new representation for interactive video generation and manipulation in space and time. With a single image at inference time, our novel framework allows the user to move objects in 3D while generating a video by providing a sequence of desired actions. The actions are learnt in an unsupervised manner. The camera can be controlled to get the desired viewpoint. Our method builds an environment state for each frame, which can be manipulated by our proposed action module and decoded back to the image space with volumetric rendering. To support diverse appearances of objects, we extend neural radiance fields with style-based modulation. Our method trains on a collection of various monocular videos requiring only the estimated camera parameters and 2D object locations. To set a challenging benchmark, we introduce two large scale video datasets with significant camera movements. As evidenced by our experiments, playable environments enable several creative applications not attainable by prior video synthesis works, including playable 3D video generation, stylization and manipulation. We make our code, demos and datasets publicly available.

****

Robust Cross-Modal Representation Learning With Progressive Self-Distillation
Alex Andonian, Shixing Chen, Raffay Hamid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16430-16441
The learning objective of vision-language approach of CLIP does not effectively account for the noisy many-to-many correspondences found in web-harvested image captioning datasets, which contributes to its compute and data inefficiency. To address this challenge, we introduce a novel training framework based on cross-modal contrastive learning that uses progressive self-distillation and soft image-text alignments to more efficiently learn robust representations from noisy data. Our model distills its own knowledge to dynamically generate soft-alignment targets for a subset of images and captions in every minibatch, which are then us

ed to update its parameters. Extensive evaluation across 14 benchmark datasets s hows that our method consistently outperforms its CLIP counterpart in multiple s ettings, including: (a) zero-shot classification, (b) linear probe transfer, and (c) image-text retrieval, without incurring added computational cost. Analysis using an ImageNet-based robustness test-bedreveals that our method offers better effective robustness to natural distribution shifts compared to both ImageNet-t rained models and CLIP itself. Lastly, pretraining with datasets spanning two or ders of magnitude in size shows that our improvements over CLIP tend to scale wi th number of training examples.

****

What To Look at and Where: Semantic and Spatial Refined Transformer for Detectin g Human-Object Interactions

A S M Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, Davide Modolo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2022, pp. 5353-5363

We propose a novel one-stage Transformer-based semantic and spatial refined tran sformer (SSRT) to solve the Human-Object Interaction detection task, which requi res to localize humans and objects, and predicts their interactions. Differently from previous Transformer-based HOI approaches, which mostly focus at improving the design of the decoder outputs for the final detection, SSRT introduces two new modules to help select the most relevant object-action pairs within an image and refine the queries' representation using rich semantic and spatial features . These enhancements lead to state-of-the-art results on the two most popular HO I benchmarks: V-COCO and HICO-DET.

****

Compressive Single-Photon 3D Cameras

Felipe Gutierrez-Barragan, Atul Ingle, Trevor Seets, Mohit Gupta, Andreas Velten ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 17854-17864

Single-photon avalanche diodes (SPADs) are an emerging pixel technology for time -of-flight (ToF) 3D cameras that can capture the time-of-arrival of individual p hotons at picosecond resolution. To estimate depths, current SPAD-based 3D camer as measure the round-trip time of a laser pulse by building a per-pixel histogra m of photon timestamps. As the spatial and timestamp resolution of SPAD-based ca meras increase, their output data rates far exceed the capacity of existing data transfer technologies. One major reason for SPAD's bandwidth-intensive operatio n is the tight coupling that exists between depth resolution and histogram resol ution. To weaken this coupling, we propose compressive single-photon histograms (CSPH). CSPHs are a per-pixel compressive representation of the high-resolution histogram, that is built on-the-fly, as each photon is detected. They are based on a family of linear coding schemes that can be expressed as a simple matrix op eration. We design different CSPH coding schemes for 3D imaging and evaluate the m under different signal and background levels, laser waveforms, and illuminatio n setups. Our results show that a well-designed CSPH can consistently reduce dat a rates by 1-2 orders of magnitude without compromising depth precision.

****

Stereo Magnification With Multi-Layer Images

Taras Khakhulin, Denis Korzhenkov, Pavel Solovev, Gleb Sterkin, Andrei-Timotei A rdelean, Victor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 8687-8696

Representing scenes with multiple semitransparent colored layers has been a popu lar and successful choice for real-time novel view synthesis. Existing approache s infer colors and transparency values over regularly spaced layers of planar or spherical shape. In this work, we introduce a new view synthesis approach based on multiple semitransparent layers with scene-adapted geometry. Our approach in fers such representations from stereo pairs in two stages. The first stage produ ces the geometry of a small number of data-adaptive layers from a given pair of views. The second stage infers the color and transparency values for these layer s, producing the final representation for novel view synthesis. Importantly, bot h stages are connected through a differentiable renderer and are trained end-to-

end. In the experiments, we demonstrate the advantage of the proposed approach o
ver the use of regularly spaced layers without adaptation to scene geometry. Des
pite being orders of magnitude faster during rendering, our approach also outper
forms the recently proposed IBRNet system based on implicit geometry representat
ion.
****

CO-SNE: Dimensionality Reduction and Visualization for Hyperbolic Data
Yunhui Guo, Haoran Guo, Stella X. Yu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21-30
Hyperbolic space can naturally embed hierarchies that often exist in real-world
data and semantics. While high dimensional hyperbolic embeddings lead to better
representations, most hyperbolic models utilize low-dimensional embeddings, due
to non-trivial optimization and visualization of high-dimensional hyperbolic dat
a. We propose CO-SNE, which extends the Euclidean space visualization tool, t-SN
E, to hyperbolic space. Like t-SNE, it converts distances between data points to
 joint probabilities and tries to minimize the Kullback-Leibler divergence betwe
en the joint probabilities of high-dimensional data X and low-dimensional embedd
ing Y. However, unlike Euclidean space, hyperbolic space is inhomogeneous: A vol
ume could contain a lot more points at a location far from the origin. CO-SNE th
us uses hyperbolic normal distributions for X and hyperbolic Cauchy instead of t
-SNE's Student's t-distribution for Y , and it additionally seeks to preserve X'
s individual distances to the Origin in Y. We apply CO-SNE to naturally hyperbol
ic data and supervisedly learned hyperbolic features. Our results demonstrate th
at CO-SNE deflates high-dimensional hyperbolic data into a low-dimensional space
 without losing their hyperbolic characteristics, significantly outperforming po
pular visualization tools such as PCA, t-SNE, UMAP, and HoroPCA which is also de
signed for hyperbolic data
****

Revisiting Skeleton-Based Action Recognition
Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, Bo Dai; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2969-29
78
Human skeleton, as a compact representation of human action, has received increa
sing attention in recent years. Many skeleton-based action recognition methods a
dopt GCNs to extract features on top of human skeletons. Despite the positive re
sults shown in these attempts, GCN-based methods are subject to limitations in r
obustness, interoperability, and scalability. In this work, we propose PoseConv3
D, a new approach to skeleton-based action recognition. PoseConv3D relies on a 3
D heatmap volume instead of a graph sequence as the base representation of human
 skeletons. Compared to GCN-based methods, PoseConv3D is more effective in learn
ing spatiotemporal features, more robust against pose estimation noises, and gen
eralizes better in cross-dataset settings. Also, PoseConv3D can handle multiple-
person scenarios without additional computation costs. The hierarchical features
 can be easily integrated with other modalities at early fusion stages, providin
g a great design space to boost the performance. PoseConv3D achieves the state-o
f-the-art on five of six standard skeleton-based action recognition benchmarks.
Once fused with other modalities, it achieves the state-of-the-art on all eight
multi-modality action recognition benchmarks. Code has been made available at: h
ttps://github.com/kennymckormick/pyskl.
****

Rethinking Controllable Variational Autoencoders
Huajie Shao, Yifei Yang, Haohong Lin, Longzhong Lin, Yizhuo Chen, Qinmin Yang, H
an Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2022, pp. 19250-19259
The Controllable Variational Autoencoder (ControlVAE) combines automatic control
 theory with the basic VAE model to manipulate the KL-divergence for overcoming
posterior collapse and learning disentangled representations. It has shown succe
ss in a variety of applications, such as image generation, disentangled represen
tation learning, and language modeling. However, when it comes to disentangled r
epresentation learning, ControlVAE does not delve into the rationale behind it.

The goal of this paper is to develop a deeper understanding of ControlVAE in lea
rning disentangled representations, including the choice of a desired KL-diverge
nce (i.e, set point), and its stability during training. We first fundamentally
explain its ability to disentangle latent variables from an information bottlene
ck perspective. We show that KL-divergence is an upper bound of the variational
information bottleneck. By controlling the KL-divergence gradually from a small
value to a target value, ControlVAE can disentangle the latent factors one by on
e. Based on this finding, we propose a new DynamicVAE that leverages a modified
incremental PI (proportional-integral) controller, a variant of the proportional
-integral-derivative (PID) algorithm, and employs a moving average as well as a
hybrid annealing method to evolve the value of KL-divergence smoothly in a tight
ly controlled fashion. In addition, we analytically derive a lower bound of the
set point for disentangling. We then theoretically prove the stability of the pr
oposed approach. Evaluation results on multiple benchmark datasets demonstrate t
hat DynamicVAE achieves a good trade-off between the disentanglement and reconst
ruction quality. We also discover that it can separate disentangled representati
on learning and reconstruction via manipulating the desired KL-divergence.
****

Contextual Instance Decoupling for Robust Multi-Person Pose Estimation
Dongkai Wang, Shiliang Zhang; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2022, pp. 11060-11068
Crowded scenes make it challenging to differentiate persons and locate their pos
e keypoints. This paper proposes the Contextual Instance Decoupling (CID), which
 presents a new pipeline for multi-person pose estimation. Instead of relying on
 person bounding boxes to spatially differentiate persons, CID decouples persons
 in an image into multiple instance-aware feature maps. Each of those feature ma
ps is hence adopted to infer keypoints for a specific person. Compared with boun
ding box detection, CID is differentiable and robust to detection errors. Decoup
ling persons into different feature maps allows to isolate distractions from oth
er persons, and explore context cues at scales larger than the bounding box size
. Experiments show that CID outperforms previous multi-person pose estimation pi
pelines on crowded scenes pose estimation benchmarks in both accuracy and effici
ency. For instance, it achieves 71.3% AP on CrowdPose, outperforming the recent
single-stage DEKR by 5.6%, the bottom-up CenterAttention by 3.7%, and the top-do
wn JC-SPPE by 5.3%. This advantage sustains on the commonly used COCO benchmark.
****

LMGP: Lifted Multicut Meets Geometry Projections for Multi-Camera Multi-Object T
racking
Duy M. H. Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, Paul Swoboda
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 8866-8875
Multi-Camera Multi-Object Tracking is currently drawing attention in the compute
r vision field due to its superior performance in real-world applications such a
s video surveillance with crowded scenes or in wide spaces. In this work, we pro
pose a mathematically elegant multi-camera multiple object tracking approach bas
ed on a spatial-temporal lifted multicut formulation. Our model utilizes state-o
f-the-art tracklets produced by single-camera trackers as proposals. As these tr
acklets may contain ID-Switch errors, we refine them through a novel pre-cluster
ing obtained from 3D geometry projections. As a result, we derive a better track
ing graph without ID switches and more precise affinity costs for the data assoc
iation phase. Tracklets are then matched to multi-camera trajectories by solving
 a global lifted multicut formulation that incorporates short and long-range tem
poral interactions on tracklets located in the same camera as well as inter-came
ra ones. Experimental results on the WildTrack dataset yield near-perfect perfor
mance, outperforming state-of-the-art trackers on Campus while being on par on t
he PETS-09 dataset.
****

Boosting Crowd Counting via Multifaceted Attention
Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, Xiaopeng Hong; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp

. 19628-19637
This paper focuses on crowd counting. As large-scale variations often exist with in crowd images, neither fixed-size convolution kernel of CNN nor fixed-size att entions of recent vision transformers can well handle this kind of variations. T o address this problem, we propose a Multifaceted Attention Network (MAN), which incorporates global attention from vanilla transformer, learnable local attenti on, attention regularization and instance attention into a counting model. First ly, the local Learnable Region Attention (LRA) is proposed to assign attention e xclusive for each feature location dynamically. Secondly, we design the Local At tention Regularization to supervise the training of LRA by minimizing the deviat ion among the attention for different feature locations. Finally, we provide an Instance Attention mechanism to focus on the most important instances dynamicall y during training. Extensive experiments on four challenging crowd counting data sets namely ShanghaiTech, UCF-QNRF, JHU++, and NWPU have validated the proposed method.
****

Stereo Depth From Events Cameras: Concentrate and Focus on the Future
Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, Jonghyun Choi; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 6114-6123
Neuromorphic cameras or event cameras mimic human vision by reporting changes in the intensity in a scene, instead of reporting the whole scene at once in a for m of an image frame as performed by conventional cameras. Events are streamed da ta that are often dense when either the scene changes or the camera moves rapidl y. The rapid movement causes the events to be overridden or missed when creating a tensor for the machine to learn on. To alleviate the event missing or overrid ing issue, we propose to learn to concentrate on the dense events to produce a c ompact event representation with high details for depth estimation. Specifically , we learn a model with events from both past and future but infer only with pas t data with the predicted future. We initially estimate depth in an event-only s etting but also propose to further incorporate images and events by a hierarchic al event and intensity combination network for better depth estimation. By exper iments in challenging real-world scenarios, we validate that our method outperfo rms prior arts even with low computational cost. Code is available at: https://g ithub.com/yonseivnl/se-cff.
****

A Probabilistic Graphical Model Based on Neural-Symbolic Reasoning for Visual Re lationship Detection
Dongran Yu, Bo Yang, Qianhao Wei, Anchen Li, Shirui Pan; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 106 09-10618
This paper aims to leverage symbolic knowledge to improve the performance and in terpretability of the Visual Relationship Detection (VRD) models. Existing VRD m ethods based on deep learning suffer from the problems of poor performance on in sufficient labeled examples and lack of interpretability. To overcome the aforem entioned weaknesses, we integrate symbolic knowledge into deep learning models a nd propose a bi-level probabilistic graphical reasoning framework called BPGR. S pecifically, in the high-level structure, we take the objects and relationships detected by the VRD model as hidden variables (reasoning results); In the low-le vel structure of BPGR, we use Markov Logic Networks (MLNs) to project First-Orde r Logic (FOL) as observed variables (symbolic knowledge) to correct error reason ing results. We adopt a variational EM algorithm for optimization. Experiments r esults show that our BPGR improves the performance of the VRD models. In particu lar, BPGR can also provide easy-to-understand insights for reasoning results to show interpretability.
****

A Simple Data Mixing Prior for Improving Self-Supervised Learning
Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, Cih ang Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 14595-14604

Data mixing (e.g., Mixup, Cutmix, ResizeMix) is an essential component for advancing recognition models. In this paper, we focus on studying its effectiveness in the self-supervised setting. By noticing the mixed images that share the same source images are intrinsically related to each other, we hereby propose SDMP, short for Simple Data Mixing Prior, to capture this straightforward yet essential prior, and position such mixed images as additional positive pairs to facilitate self-supervised representation learning. Our experiments verify that the proposed SDMP enables data mixing to help a set of self-supervised learning frameworks (e.g., MoCo) achieve better accuracy and out-of-distribution robustness. More notably, our SDMP is the first method that successfully leverages data mixing to improve (rather than hurt) the performance of Vision Transformers in the self-supervised setting. Code is publicly available at https://github.com/OliverRensu/SDMP.

****

Knowledge Distillation As Efficient Pre-Training: Faster Convergence, Higher Data-Efficiency, and Better Transferability

Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9161-9171

Large-scale pre-training has been proven to be crucial for various computer vision tasks. However, with the increase of pre-training data amount, model architecture amount, and the private/inaccessible data, it is not very efficient or possible to pre-train all the model architectures on large-scale datasets. In this work, we investigate an alternative strategy for pre-training, namely Knowledge Distillation as Efficient Pre-training (KDEP), aiming to efficiently transfer the learned feature representation from existing pre-trained models to new student models for future downstream tasks. We observe that existing Knowledge Distillation (KD) methods are unsuitable towards pre-training since they normally distill the logits that are going to be discarded when transferred to downstream tasks. To resolve this problem, we propose a feature-based KD method with non-parametric feature dimension aligning. Notably, our method performs comparably with supervised pre-training counterparts in 3 downstream tasks and 9 downstream datasets requiring 10x less data and 5x less pre-training time. Code is available at https://github.com/CVMI-Lab/KDEP.

****

LOLNerf: Learn From One Look

Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1558-1567

We present a method for learning a generative 3D model based on neural radiance fields, trained solely from data with only single views of each object. While generating realistic images is no longer a difficult task, producing the corresponding 3D structure such that they can be rendered from different views is non-trivial. We show that, unlike existing methods, one does not need multi-view data to achieve this goal. Specifically, we show that by reconstructing many images aligned to an approximate canonical pose with a single network conditioned on a shared latent space, you can learn a space of radiance fields that models shape and appearance for a class of objects. We demonstrate this by training models to reconstruct object categories using datasets that contain only one view of each subject without depth or geometry information. Our experiments show that we achieve state-of-the-art results in novel view synthesis and high-quality results for monocular depth prediction.

****

Geometry-Aware Guided Loss for Deep Crack Recognition

Zhuangzhuang Chen, Jin Zhang, Zhuonan Lai, Jie Chen, Zun Liu, Jianqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4703-4712

Despite the substantial progress of deep models for crack recognition, due to the inconsistent cracks in varying sizes, shapes, and noisy background textures, there still lacks the discriminative power of the deeply learned features when su

pervised by the cross-entropy loss. In this paper, we propose the geometry-aware guided loss (GAGL) that enhances the discrimination ability and is only applied in the training stage without extra computation and memory during inference. The GAGL consists of the feature-based geometry-aware projected gradient descent method (FGA-PGD) that approximates the geometric distances of the features to the class boundaries, and the geometry-aware update rule that learns an anchor of each class as the approximation of the feature expected to have the largest geometric distance to the corresponding class boundary. Then the discriminative power can be enhanced by minimizing the distances between the features and their corresponding class anchors in the feature space. To address the limited availability of related benchmarks, we collect a fully annotated dataset, namely, NPP2021, which involves inconsistent cracks and noisy backgrounds in real-world nuclear power plants. Our proposed GAGL outperforms the state of the arts on various benchmark datasets including CRACK2019, SDNET2018, and our NPP2021.

****

Multi-Modal Alignment Using Representation Codebook

Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, Trishul Chilimbi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15651-15660

Aligning signals from different modalities is an important step in vision-language representation learning as it affects the performance of later stages such as cross-modality fusion. Since image and text typically reside in different regions of the feature space, directly aligning them at instance level is challenging especially when features are still evolving during training. In this paper, we propose to align at a higher and more stable level using cluster representation. Specifically, we treat image and text as two views of the same entity, and encode them into a joint vision-language coding space spanned by a dictionary of cluster centers (codebook). We contrast positive and negative samples via their cluster assignments while simultaneously optimizing the cluster centers. To further smooth out the learning process, we adopt a teacher-student distillation paradigm, where the momentum teacher of one view guides the student learning of the other. We evaluated our approach on common vision language benchmarks and obtain new SoTA on zero-shot cross modality retrieval while being competitive on various other transfer tasks.

****

Maintaining Reasoning Consistency in Compositional Visual Question Answering

Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5099-5108

A compositional question refers to a question that contains multiple visual concepts (e.g., objects, attributes, and relationships) and requires compositional reasoning to answer. Existing VQA models can answer a compositional question well, but cannot work well in terms of reasoning consistency in answering the compositional question and its sub-questions. For example, a compositional question for an image is: "Are there any elephants to the right of the white bird?" and one of its sub-questions is " Is any bird visible in the scene?". The models may answer "yes" to the compositional question, but "no" to the sub-question. This paper presents a dialog-like reasoning method for maintaining reasoning consistency in answering a compositional question and its sub-questions. Our method integrates the reasoning processes for the sub-questions into the reasoning process for the compositional question like a dialog task, and uses a consistency constraint to penalize inconsistent answer predictions. In order to enable quantitative evaluation of reasoning consistency, we construct a GQA-Sub dataset based on the well-organized GQA dataset. Experimental results on the GQA dataset and the GQA-Sub dataset demonstrate the effectiveness of our method.

****

Structure-Aware Motion Transfer With Deformable Anchor Model

Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, Lixin Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3637-3646

Given a source image and a driving video depicting the same object type, the mot ion transfer task aims to generate a video by learning the motion from the drivi ng video while preserving the appearance from the source image. In this paper, w e propose a novel structure-aware motion modeling approach, the deformable ancho r model (DAM), which can automatically discover the motion structure of arbitrar y objects without leveraging their prior structure information. Specifically, in spired by the known deformable part model (DPM), our DAM introduces two types of anchors or keypoints: i) a number of motion anchors that capture both appearanc e and motion information from the source image and driving video; ii) a latent r oot anchor, which is linked to the motion anchors to facilitate better learning of the representations of the object structure information. Moreover, DAM can be further extended to a hierarchical version through the introduction of addition al latent anchors to model more complicated structures. By regularizing motion a nchors with latent anchor(s), DAM enforces the correspondences between them to e nsure the structural information is well captured and preserved. Moreover, DAM c an be learned effectively in an unsupervised manner. We validate our proposed DA M for motion transfer on different benchmark datasets. Extensive experiments cle arly demonstrate that DAM achieves superior performance relative to existing sta te-of-the-art methods.

****

BigDL 2.0: Seamless Scaling of AI Pipelines From Laptops to Distributed Cluster

Jason (Jinquan) Dai, Ding Ding, Dongjie Shi, Shengsheng Huang, Jiao Wang, Xin Qi u, Kai Huang, Guoqiong Song, Yang Wang, Qiyuan Gong, Jiaming Song, Shan Yu, Le Z heng, Yina Chen, Junwei Deng, Ge Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21439-21446

Most AI projects start with a Python notebook running on a single laptop; howeve r, one usually needs to go through a mountain of pains to scale it to handle lar ger dataset (for both experimentation and production deployment). These usually entail many manual and error-prone steps for the data scientists to fully take a dvantage of the available hardware resources (e.g., SIMD instructions, multi-pro cessing, quantization, memory allocation optimization, data partitioning, distri buted computing, etc.).To address this challenge, we have open sourced BigDL 2.0 at https://github.com/intel-analytics/BigDL/ under Apache 2.0 license (combinin g the original BigDL [19] and Analytics Zoo [18] projects); using BigDL 2.0, use rs can simply build conventional Python notebooks on their laptops (with possibl e AutoML support), which can then be transparently accelerated on a single node (with up-to 9.6x speedup in our experiments), and seamlessly scaled out to a lar ge cluster (across several hundreds servers in real-world use cases). BigDL 2.0 has already been adopted by many real-world users (such as Mastercard, Burger Ki ng, Inspur, etc.) in production.

****

Integrative Few-Shot Learning for Classification and Segmentation

Dahyun Kang, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 9979-9990

We introduce the integrative task of few-shot classification and segmentation (F S-CS) that aims to both classify and segment target objects in a query image whe n the target classes are given with a few examples. This task combines two conve ntional few-shot learning problems, few-shot classification and segmentation. FS -CS generalizes them to more realistic episodes with arbitrary image pairs, wher e each target class may or may not be present in the query. To address the task, we propose the integrative few-shot learning (iFSL) framework for FS-CS, which trains a learner to construct class-wise foreground maps for multi-label classif ication and pixel-wise segmentation. We also develop an effective iFSL model, at tentive squeeze network (ASNet), that leverages deep semantic correlation and gl obal self-attention to produce reliable foreground maps. In experiments, the pro posed method shows promising performance on the FS-CS task and also achieves the state of the art on standard few-shot segmentation benchmarks

****

Acquiring a Dynamic Light Field Through a Single-Shot Coded Image

Ryoya Mizuno, Keita Takahashi, Michitaka Yoshida, Chihiro Tsutake, Toshiaki Fuji

i, Hajime Nagahara; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2022, pp. 19830-19840
We propose a method for compressively acquiring a dynamic light field (a 5-D vol
ume) through a single-shot coded image (a 2-D measurement). We designed an imagi
ng model that synchronously applies aperture coding and pixel-wise exposure codi
ng within a single exposure time. This coding scheme enables us to effectively e
mbed the original information into a single observed image. The observed image i
s then fed to a convolutional neural network (CNN) for light-field reconstructio
n, which is jointly trained with the camera-side coding patterns. We also develo
ped a hardware prototype to capture a real 3-D scene moving over time. We succee
ded in acquiring a dynamic light field with 5x5 viewpoints over 4 temporal sub-f
rames (100 views in total) from a single observed image. Repeating capture and r
econstruction processes over time, we can acquire a dynamic light field at 4x th
e frame rate of the camera. To our knowledge, our method is the first to achieve
 a finer temporal resolution than the camera itself in compressive light-field a
cquisition. Our software is available from our project webpage.
****

Attentive Fine-Grained Structured Sparsity for Image Restoration
Junghun Oh, Heewon Kim, Seungjun Nah, Cheeun Hong, Jonghyun Choi, Kyoung Mu Lee;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2022, pp. 17673-17682
Image restoration tasks have witnessed great performance improvement in recent y
ears by developing large deep models. Despite the outstanding performance, the h
eavy computation demanded by the deep models has restricted the application of i
mage restoration. To lift the restriction, it is required to reduce the size of
the networks while maintaining accuracy. Recently, N:M structured pruning has ap
peared as one of the effective and practical pruning approaches for making the m
odel efficient with the accuracy constraint. However, it fails to account for di
fferent computational complexities and performance requirements for different la
yers of an image restoration network. To further optimize the trade-off between
the efficiency and the restoration accuracy, we propose a novel pruning method t
hat determines the pruning ratio for N:M structured sparsity at each layer. Exte
nsive experimental results on super-resolution and deblurring tasks demonstrate
the efficacy of our method which outperforms previous pruning methods significan
tly. PyTorch implementation for the proposed methods will be publicly available
at https://github.com/JungHunOh/SLS_CVPR2022
****

Pix2NeRF: Unsupervised Conditional p-GAN for Single Image to Neural Radiance Fie
lds Translation
Shengqu Cai, Anton Obukhov, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3981-
3990
We propose a pipeline to generate Neural Radiance Fields (NeRF) of an object or
a scene of a specific class, conditioned on a single input image. This is a chal
lenging task, as training NeRF requires multiple views of the same scene, couple
d with corresponding poses, which are hard to obtain. Our method is based on pi-
GAN, a generative model for unconditional 3D-aware image synthesis, which maps r
andom latent codes to radiance fields of a class of objects. We jointly optimize
 (1) the pi-GAN objective to utilize its high-fidelity 3D-aware generation and (
2) a carefully designed reconstruction objective. The latter includes an encoder
 coupled with pi-GAN generator to form an auto-encoder. Unlike previous few-shot
 NeRF approaches, our pipeline is unsupervised, capable of being trained with in
dependent images without 3D, multi-view, or pose supervision. Applications of ou
r pipeline include 3d avatar generation, object-centric novel view synthesis wit
h a single input image, and 3d-aware super-resolution, to name a few.
****

HARA: A Hierarchical Approach for Robust Rotation Averaging
Seong Hun Lee, Javier Civera; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2022, pp. 15777-15786
We propose a novel hierarchical approach for multiple rotation averaging, dubbed

HARA. Our method incrementally initializes the rotation graph based on a hierarchy of triplet support. The key idea is to build a spanning tree by prioritizing the edges with many strong triplet supports and gradually adding those with weaker and fewer supports. This reduces the risk of adding outliers in the spanning tree. As a result, we obtain a robust initial solution that enables us to filter outliers prior to nonlinear optimization. With minimal modification, our approach can also integrate the knowledge of the number of valid 2D-2D correspondences. We perform extensive evaluations on both synthetic and real datasets, demonstrating state-of-the-art results.
****

Diffusion Autoencoders: Toward a Meaningful and Decodable Representation
Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10619-10629
Diffusion probabilistic models (DPMs) have achieved remarkable quality in image generation that rivals GANs'. But unlike GANs, DPMs use a set of latent variables that lack semantic meaning and cannot serve as a useful representation for other tasks. This paper explores the possibility of using DPMs for representation learning and seeks to extract a meaningful and decodable representation of an input image via autoencoding. Our key idea is to use a learnable encoder for discovering the high-level semantics, and a DPM as the decoder for modeling the remaining stochastic variations. Our method can encode any image into a two-part latent code, where the first part is semantically meaningful and linear, and the second part captures stochastic details, allowing near-exact reconstruction. This capability enables challenging applications that currently foil GAN-based methods, such as attribute manipulation on real images. We also show that this two-level encoding improves denoising efficiency and naturally facilitates various downstream tasks including few-shot conditional sampling.
****

Learning Fair Classifiers With Partially Annotated Group Labels
Sangwon Jung, Sanghyuk Chun, Taesup Moon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10348-10357
Recently, fairness-aware learning have become increasingly crucial, but most of those methods operate by assuming the availability of fully annotated demographic group labels. We emphasize that such assumption is unrealistic for real-world applications since group label annotations are expensive and can conflict with privacy issues. In this paper, we consider a more practical scenario, dubbed as Algorithmic Group Fairness with the Partially annotated Group labels (Fair-PG). We observe that the existing methods to achieve group fairness perform even worse than the vanilla training, which simply uses full data only with target labels, under Fair-PG. To address this problem, we propose a simple Confidence-based Group Label assignment (CGL) strategy that is readily applicable to any fairness-aware learning method. CGL utilizes an auxiliary group classifier to assign pseudo group labels, where random labels are assigned to low confident samples. We first theoretically show that our method design is better than the vanilla pseudo-labeling strategy in terms of fairness criteria. Then, we empirically show on several benchmark datasets that by combining CGL and the state-of-the-art fairness-aware in-processing methods, the target accuracies and the fairness metrics can be jointly improved compared to the baselines. Furthermore, we convincingly show that CGL enables to naturally augment the given group-labeled dataset with external target label-only datasets so that both accuracy and fairness can be improved. Code is available at https: //github.com/naver-ai/cgl_fairness.
****

StylizedNeRF: Consistent 3D Scene Stylization As Stylized NeRF via 2D-3D Mutual Learning
Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, Lin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18342-18352
3D scene stylization aims at generating stylized images of the scene from arbitrary novel views following a given set of style examples, while ensuring consiste

ncy when rendered from different views. Directly applying methods for image or v ideo stylization to 3D scenes cannot achieve such consistency. Thanks to recentl y proposed neural radiance fields (NeRF), we are able to represent a 3D scene in a consistent way. Consistent 3D scene stylization can be effectively achieved b y stylizing the corresponding NeRF. However, there is a significant domain gap b etween style examples which are 2D images and NeRF which is an implicit volumetr ic representation. To address this problem, we propose a novel mutual learning f ramework for 3D scene stylization that combines a 2D image stylization network a nd NeRF to fuse the stylization ability of 2D stylization network with the 3D co nsistency of NeRF. We first pre-train a standard NeRF of the 3D scene to be styl ized and replace its color prediction module with a style network to obtain a st ylized NeRF. It is followed by distilling the prior knowledge of spatial consist ency from NeRF to the 2D stylization network through an introduced consistency l oss. We also introduce a mimic loss to supervise the mutual learning of the NeRF style module and fine-tune the 2D stylization decoder. In order to further make our model handle ambiguities of 2D stylization results, we introduce learnable latent codes that obey the probability distributions conditioned on the style. T hey are attached to training samples as conditional inputs to better learn the s tyle module in our novel stylized NeRF. Experimental results demonstrate that ou r method is superior to existing approaches in both visual quality and long-rang e consistency.

****

NightLab: A Dual-Level Architecture With Hardness Detection for Segmentation at Night

Xueqing Deng, Peng Wang, Xiaochen Lian, Shawn Newsam; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16938-16948

The semantic segmentation of nighttime scenes is a challenging problem that is k ey to impactful applications like self-driving cars. Yet, it has received little attention compared to its daytime counterpart. In this paper, we propose NightL ab, a novel nighttime segmentation framework that leverages multiple deep learni ng models imbued with night-aware features to yield State-of-The-Art (SoTA) perf ormance on multiple night segmentation benchmarks. Notably, NightLab contains mo dels at two levels of granularity, i.e. image and regional, and each level is co mposed of light adaptation and segmentation modules. Given a nighttime image, th e image level model provides an initial segmentation estimate while, in parallel , a hardness detection module identifies regions and their surrounding context t hat need further analysis. A regional level model focuses on these difficult reg ions to provide a significantly improved segmentation. All the models in NightLa b are trained end-to-end using a set of proposed night-aware losses without hand crafted heuristics. Extensive experiments on the NightCity and BDD100K datasets show NightLab achieves SoTA performance compared to concurrent methods. Code and dataset are available at https://github.com/xdeng7/NightLab.

****

Knowledge Distillation With the Reused Teacher Classifier

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, Chun Chen; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11933-11942

Knowledge distillation aims to compress a powerful yet cumbersome teacher model into a lightweight student model without much sacrifice of performance. For this purpose, various approaches have been proposed over the past few years, general ly with elaborately designed knowledge representations, which in turn increase t he difficulty of model development and interpretation. In contrast, we empirical ly show that a simple knowledge distillation technique is enough to significantl y narrow down the teacher-student performance gap. We directly reuse the discrim inative classifier from the pre-trained teacher model for student inference and train a student encoder through feature alignment with a single L2 loss. In this way, the student model is able to achieve exactly the same performance as the t eacher model provided that their extracted features are perfectly aligned. An ad ditional projector is developed to help the student encoder match with the teach

er classifier, which renders our technique applicable to various teacher and stu
dent architectures. Extensive experiments demonstrate that our technique achieve
s state-of-the-art results at the modest cost of compression ratio due to the ad
ded projector.
****

Contrastive Learning for Unsupervised Video Highlight Detection
Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, Li Cheng; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14
042-14052
Video highlight detection can greatly simplify video browsing, potentially pavin
g the way for a wide range of applications. Existing efforts are mostly fully-su
pervised, requiring humans to manually identify and label the interesting moment
s (called highlights) in a video. Recent weakly supervised methods forgo the use
 of highlight annotations, but typically require extensive efforts in collecting
 external data such as web-crawled videos for model learning. This observation h
as inspired us to consider unsupervised highlight detection where neither frame-
level nor video-level annotations are available in training. We propose a simple
 contrastive learning framework for unsupervised highlight detection. Our framew
ork encodes a video into a vector representation by learning to pick video clips
 that help to distinguish it from other videos via a contrastive objective using
 dropout noise. This inherently allows our framework to identify video clips cor
responding to highlight of the video. Extensive empirical evaluations on three h
ighlight detection benchmarks demonstrate the superior performance of our approa
ch.
****

InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition
Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, Karthi
k Ramani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2022, pp. 20186-20196
Human skeleton-based action recognition offers a valuable means to understand th
e intricacies of human behavior because it can handle the complex relationships
between physical constraints and intention. Although several studies have focuse
d on encoding a skeleton, less attention has been paid to embed this information
 into the latent representations of human action. InfoGCN proposes a learning fr
amework for action recognition combining a novel learning objective and an encod
ing method. First, we design an information bottleneck-based learning objective
to guide the model to learn informative but compact latent representations. To p
rovide discriminative information for classifying action, we introduce attention
-based graph convolution that captures the context-dependent intrinsic topology
of human action. In addition, we present a multi-modal representation of the ske
leton using the relative position of joints, designed to provide complementary s
patial information for joints. InfoGCN surpasses the known state-of-the-art on m
ultiple skeleton-based action recognition benchmarks with the accuracy of 93.0%
on NTU RGB+D 60 cross-subject split, 89.8% on NTU RGB+D 120 cross-subject split,
 and 97.0% on NW-UCLA.
****

Rethinking Image Cropping: Exploring Diverse Compositions From Global Views
Gengyun Jia, Huaibo Huang, Chaoyou Fu, Ran He; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2446-2455
Existing image cropping works mainly use anchor evaluation methods or coordinate
 regression methods. However, it is difficult for pre-defined anchors to cover g
ood crops globally, and the regression methods ignore the cropping diversity. In
 this paper, we regard image cropping as a set prediction problem. A set of crop
s regressed from multiple learnable anchors is matched with the labeled good cro
ps, and a classifier is trained using the matching results to select a valid sub
set from all the predictions. This new perspective equips our model with globali
ty and diversity, mitigating the shortcomings but inherit the strengthens of pre
vious methods. Despite the advantages, the set prediction method causes inconsis
tency between the validity labels and the crops. To deal with this problem, we p
ropose to smooth the validity labels with two different methods. The first metho

d that uses crop qualities as direct guidance is designed for the datasets with nearly dense quality labels. The second method based on the self distillation can be used in sparsely labeled datasets. Experimental results on the public datasets show the merits of our approach over state-of-the-art counterparts.
****

Constrained Few-Shot Class-Incremental Learning
Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, Abbas Rahimi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9057-9067
Continually learning new classes from fresh data without forgetting previous knowledge of old classes is a very challenging research problem. Moreover, it is imperative that such learning must respect certain memory and computational constraints such as (i) training samples are limited to only a few per class, (ii) the computational cost of learning a novel class remains constant, and (iii) the memory footprint of the model grows at most linearly with the number of classes observed. To meet the above constraints, we propose C-FSCIL, which is architecturally composed of a frozen meta-learned feature extractor, a trainable fixed-size fully connected layer, and a rewritable dynamically growing memory that stores as many vectors as the number of encountered classes. C-FSCIL provides three update modes that offer a trade-off between accuracy and compute-memory cost of learning novel classes. C-FSCIL exploits hyperdimensional embedding that allows to continually express many more classes than the fixed dimensions in the vector space, with minimal interference. The quality of class vector representations is further improved by aligning them quasi-orthogonally to each other by means of novel loss functions. Experiments on the CIFAR100, miniImageNet, and Omniglot datasets show that C-FSCIL outperforms the baselines with remarkable accuracy and compression. It also scales up to the largest problem size ever tried in this few-shot setting by learning 423 novel classes on top of 1200 base classes with less than 1.6% accuracy drop. Our code is available at https://github.com/IBM/constrained-FSCIL
****

Self-Supervised Material and Texture Representation Learning for Remote Sensing Tasks
Peri Akiva, Matthew Purri, Matthew Leotta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8203-8215
Self-supervised learning aims to learn image feature representations without the usage of manually annotated labels. It is often used as a precursor step to obtain useful initial network weights which contribute to faster convergence and superior performance of downstream tasks. While self-supervision allows one to reduce the domain gap between supervised and unsupervised learning without the usage of labels, the self-supervised objective still requires a strong inductive bias to downstream tasks for effective transfer learning. In this work, we present our material and texture based self-supervision method named MATTER (MATerial and TExture Representation Learning), which is inspired by classical material and texture methods. Material and texture can effectively describe any surface, including its tactile properties, color, and specularity. By extension, effective representation of material and texture can describe other semantic classes strongly associated with said material and texture. MATTER leverages multi-temporal, spatially aligned remote sensing imagery over unchanged regions to learn invariance to illumination and viewing angle as a mechanism to achieve consistency of material and texture representation. We show that our self-supervision pre-training method allows for up to 24.22% and 6.33% performance increase in unsupervised and fine-tuned setups, and up to 76% faster convergence on change detection, land cover classification, and semantic segmentation tasks.
****

Threshold Matters in WSSS: Manipulating the Activation for the Robust and Accurate Segmentation Model Against Thresholds
Minhyun Lee, Dongseob Kim, Hyunjung Shim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4330-4339
Weakly-supervised semantic segmentation (WSSS) has recently gained much attentio

n for its promise to train segmentation models only with image-level labels. Existing WSSS methods commonly argue that the sparse coverage of CAM incurs the performance bottleneck of WSSS. This paper provides analytical and empirical evidence that the actual bottleneck may not be sparse coverage but a global thresholding scheme applied after CAM. Then, we show that this issue can be mitigated by satisfying two conditions; 1) reducing the imbalance in the foreground activation and 2) increasing the gap between the foreground and the background activation. Based on these findings, we propose a novel activation manipulation network with a per-pixel classification loss and a label conditioning module. Per-pixel classification naturally induces two-level activation in activation maps, which can penalize the most discriminative parts, promote the less discriminative parts, and deactivate the background regions. Label conditioning imposes that the output label of pseudo-masks should be any of true image-level labels; it penalizes the wrong activation assigned to non-target classes. Based on extensive analysis and evaluations, we demonstrate that each component helps produce accurate pseudo-masks, achieving the robustness against the choice of the global threshold. Finally, our model achieves state-of-the-art records on both PASCAL VOC 2012 and MS COCO 2014 datasets. The code is available at https://github.com/gaviotas/AMN.

****

Data-Free Network Compression via Parametric Non-Uniform Mixed Precision Quantization

Vladimir Chikin, Mikhail Antiukh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 450-459

Deep Neural Networks (DNNs) usually have a large number of parameters and consume a huge volume of storage space, which limits the application of DNNs on memory-constrained devices. Network quantization is an appealing way to compress DNNs. However, most of existing quantization methods require the training dataset and a fine-tuning procedure to preserve the quality of a full-precision model. These are unavailable for the confidential scenarios due to personal privacy and security problems. Focusing on this issue, we propose a novel data-free method for network compression called PNMQ, which employs the Parametric Non-uniform Mixed precision Quantization to generate a quantized network. During the compression stage, the optimal parametric non-uniform quantization grid is calculated directly for each layer to minimize the quantization error. User can directly specify the required compression ratio of a network, which is used by the PNMQ algorithm to select bitwidths of layers. This method does not require any model retraining or expensive calculations, which allows efficient implementations for network compression on edge devices. Extensive experiments have been conducted on various computer vision tasks and the results demonstrate that PNMQ achieves better performance than other state-of-the-art methods of network compression.

****

Sparse to Dense Dynamic 3D Facial Expression Generation

Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, Alberto Del Bimbo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20385-20394

In this paper, we propose a solution to the task of generating dynamic 3D facial expressions from a neutral 3D face and an expression label. This involves solving two sub-problems: (i) modeling the temporal dynamics of expressions, and (ii) deforming the neutral mesh to obtain the expressive counterpart. We represent the temporal evolution of expressions using the motion of a sparse set of 3D landmarks that we learn to generate by training a manifold-valued GAN (Motion3DGAN). To better encode the expression-induced deformation and disentangle it from the identity information, the generated motion is represented as per-frame displacement from a neutral configuration. To generate the expressive meshes, we train a Sparse2Dense mesh Decoder (S2D-Dec) that maps the landmark displacements to a dense, per-vertex displacement. This allows us to learn how the motion of a sparse set of landmarks influences the deformation of the overall face surface, independently from the identity. Experimental results on the CoMA and D3DFACS datasets show that our solution brings significant improvements with respect to previous solutions in terms of both dynamic expression generation and mesh reconstructi

on, while retaining good generalization to unseen data. The code and the pretrai
ned model will be made publicly available.
****

Think Twice Before Detecting GAN-Generated Fake Images From Their Spectral Domai
n Imprints
Chengdong Dong, Ajay Kumar, Eryun Liu; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7865-7874
Accurate detection of the fake but photorealistic images is one of the most chal
lenging tasks to address social, biometrics security and privacy related concern
s in our community. Earlier research has underlined the existence of spectral do
main artifacts in fake images generated by powerful generative adversarial netwo
rk (GAN) based methods. Therefore, a number of highly accurate frequency domain
 methods to detect such GAN generated images have been proposed in the literature
. Our study in this paper introduces a pipeline to mitigate the spectral artifac
ts. We show from our experiments that the artifacts in frequency spectrum of suc
h fake images can be mitigated by proposed methods, which leads to the sharp dec
rease of performance of spectrum-based detectors. This paper also presents exper
imental results using a large database of images that are synthesized using BigG
AN, CRN, CycleGAN, IMLE, ProGAN, StarGAN, StyleGAN and StyleGAN2 (including synt
hesized high resolution fingerprint images) to illustrate effectiveness of the p
roposed methods. Furthermore, we select a spatial-domain based fake image detect
or and observe a notable decrease in the detection performance when proposed met
hod is incorporated. In summary, our insightful analysis and pipeline presented
 in this paper cautions the forensic community on the reliability of GAN-generate
d fake image detectors that are based on the analysis of frequency artifacts as
 these artifacts can be easily mitigated.
****

Crafting Better Contrastive Views for Siamese Representation Learning
Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, Yang You; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1603
1-16040
Recent self-supervised contrastive learning methods greatly benefit from the Sia
mese structure that aims at minimizing distances between positive pairs. For hig
h performance Siamese representation learning, one of the keys is to design good
 contrastive pairs. Most previous works simply apply random sampling to make dif
ferent crops of the same image, which overlooks the semantic information that ma
y degrade the quality of views. In this work, we propose ContrastiveCrop, which
 could effectively generate better crops for Siamese representation learning. Fir
stly, a semantic-aware object localization strategy is proposed within the train
ing process in a fully unsupervised manner. This guides us to generate contrasti
ve views which could avoid most false positives (i.e., object v.s. background).
 Moreover, we empirically find that views with similar appearances are trivial fo
r the Siamese model training. Thus, a center-suppressed sampling is further desi
gned to enlarge the variance of crops. Remarkably, our method takes a careful co
nsideration of positive pairs for contrastive learning with negligible extra tra
ining overhead. As a plug-and-play and framework-agnostic module, ContrastiveCro
p consistently improves SimCLR, MoCo, BYOL, SimSiam by 0.4% 2.0% classificatio
n accuracy on CIFAR-10, CIFAR-100, Tiny ImageNet, and STL-10. Superior results a
re also achieved on downstream detection and segmentation tasks when pre-trained
 on ImageNet-1K.
****

RSCFed: Random Sampling Consensus Federated Semi-Supervised Learning
Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, Xiaomeng Li; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp.
10154-10163
Federated semi-supervised learning (FSSL) aims to derive a global model by joint
ly training fully-labeled and fully-unlabeled clients. The existing approaches w
ork well when local clients have independent and identically distributed (IID) d
ata but fail to generalize to a more practical FSSL setting, i.e., Non-IID setti
ng. In this paper, we present a Random Sampling Consensus Federated learning, na

mely RSCFed, by considering the uneven reliability among models from labeled cli
ents and unlabeled clients. Our key motivation is that given models with large d
eviations from either labeled clients or unlabeled clients, the consensus could
be reached by performing random sup-sampling over clients. To achieve it, instea
d of directly aggregating local models, we first distill several sub-consensus m
odels by random sub-sampling over clients and then aggregating the sub-consensus
 models to the global model. To enhance the robustness of sub-consensus models,
we also develop a novel distance-reweighted model aggregation method. Experiment
al results show that our method outperforms state-of-the-art methods on three be
nchmarked datasets, including both natural images and medical images.
****

TransMVSNet: Global Context-Aware Multi-View Stereo Network With Transformers
Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang W
ang, Xiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2022, pp. 8585-8594
In this paper, we present TransMVSNet, based on our exploration of feature match
ing in multi-view stereo (MVS). We analogize MVS back to its nature of a feature
 matching task and therefore propose a powerful Feature Matching Transformer (FM
T) to leverage intra- (self-) and inter- (cross-) attention to aggregate long-ra
nge context information within and across images. To facilitate a better adaptat
ion of the FMT, we leverage an Adaptive Receptive Field (ARF) module to ensure a
 smooth transit in scopes of features and bridge different stages with a feature
 pathway to pass transformed features and gradients across different scales. In
addition, we apply pair-wise feature correlation to measure similarity between f
eatures, and adopt ambiguity-reducing focal loss to strengthen the supervision.
To the best of our knowledge, TransMVSNet is the first attempt to leverage Trans
former into the task of MVS. As a result, our method achieves state-of-the-art p
erformance on DTU dataset, Tanks and Temples benchmark and BlendedMVS dataset. C
ode is available at https://github.com/MegviiRobot/TransMVSNet.
****

ROCA: Robust CAD Model Retrieval and Alignment From a Single Image
Can Gümeli, Angela Dai, Matthias Nießner; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4022-4031
We present ROCA, a novel end-to-end approach that retrieves and aligns 3D CAD mo
dels from a shape database to a single input image. This enables 3D perception o
f an observed scene from a 2D RGB observation, characterized as a lightweight, c
ompact, clean CAD representation. Core to our approach is our differentiable ali
gnment optimization based on dense 2D-3D object correspondences and Procrustes a
lignment. ROCA can thus provide a robust CAD alignment while simultaneously info
rming CAD retrieval by leveraging the 2D-3D correspondences to learn geometrical
ly similar CAD models. Experiments on challenging, real-world imagery from ScanN
et show that ROCA significantly improves on state of the art, from 9.5% to 17.6%
 in retrieval-aware CAD alignment accuracy.
****

Continual Learning for Visual Search With Backward Consistent Feature Embedding
Timmy S. T. Wan, Jun-Cheng Chen, Tzer-Yi Wu, Chu-Song Chen; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp.
16702-16711
In visual search, the gallery set could be incrementally growing and added to th
e database in practice. However, existing methods rely on the model trained on t
he entire dataset, ignoring the continual updating of the model. Besides, as the
 model updates, the new model must re-extract features for the entire gallery se
t to maintain compatible feature space, imposing a high computational cost for a
 large gallery set. To address the issues of long-term visual search, we introdu
ce a continual learning (CL) approach that can handle the incrementally growing
gallery set with backward embedding consistency. We enforce the losses of inter-
session data coherence, neighbor-session model coherence, and intra-session disc
rimination to conduct a continual learner. In addition to the disjoint setup, ou
r CL solution also tackles the situation of increasingly adding new classes for
the blurry boundary without assuming all categories known in the beginning and d

uring model update. To our knowledge, this is the first CL method both tackling the issue of backward-consistent feature embedding and allowing novel classes to occur in the new sessions. Extensive experiments on various benchmarks show the efficacy of our approach under a wide range of setups.
****

iFS-RCNN: An Incremental Few-Shot Instance Segmenter
Khoi Nguyen, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 7010-7019
This paper addresses incremental few-shot instance segmentation, where a few exa mples of new object classes arrive when access to training examples of old class es is not available anymore, and the goal is to perform well on both old and new classes. We make two contributions by extending the common Mask-RCNN framework in its second stage -- namely, we specify a new object class classifier based on the probit function and a new uncertainty-guided bounding-box predictor. The fo rmer leverages Bayesian learning to address a paucity of training examples of ne w classes. The latter learns not only to predict object bounding boxes but also to estimate the uncertainty of the prediction as a guidance for bounding box ref inement. We also specify two new loss functions in terms of the estimated object -class distribution and bounding-box uncertainty. Our contributions produce sign ificant performance gains on the COCO dataset over the state of the art -- speci fically, the gain of +6 on the new classes and +16 on the old classes in the AP instance segmentation metric. Furthermore, we are the first to evaluate the incr emental few-shot setting on the more challenging LVIS dataset.
****

DPGEN: Differentially Private Generative Energy-Guided Network for Natural Image Synthesis
Jia-Wei Chen, Chia-Mu Yu, Ching-Chia Kao, Tzai-Wei Pang, Chun-Shien Lu; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2022, pp. 8387-8396
Despite an increased demand for valuable data, the privacy concerns associated w ith sensitive datasets present a barrier to data sharing. One may use differenti ally private generative models to generate synthetic data. Unfortunately, genera tors are typically restricted to generating images of low-resolutions due to the limitation of noisy gradients. Here, we propose DPGEN, a network model designed to synthesize high-resolution natural images while satisfying differential priv acy. In particular, we propose an energy-guided network trained on sanitized dat a to indicate the direction of the true data distribution via Langevin Markov ch ain Monte Carlo (MCMC) sampling method. In contrast to the state-of-the-art meth ods that can process only low-resolution images (e.g., MNIST and Fashion-MNIST), DPGEN can generate differentially private synthetic images with resolutions up to 128*128 with superior visual quality and data utility. Our code is available at https://github.com/chiamuyu/DPGEN
****

MetaFSCIL: A Meta-Learning Approach for Few-Shot Class Incremental Learning
Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, Jin Tang; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14166-14175
In this paper, we tackle the problem of few-shot class incremental learning (FSC IL). FSCIL aims to incrementally learn new classes with only a few samples in ea ch class. Most existing methods only consider the incremental steps at test time . The learning objective of these methods is often hand-engineered and is not di rectly tied to the objective (i.e. incrementally learning new classes) during te sting. Those methods are sub-optimal due to the misalignment between the trainin g objectives and what the methods are expected to do during evaluation. In this work, we proposed a bi-level optimization based on meta-learning to directly opt imize the network to learn how to incrementally learn in the setting of FSCIL. C oncretely, we propose to sample sequences of incremental tasks from base classes for training to simulate the evaluation protocol. For each task, the model is l earned using a meta-objective such that it is capable to perform fast adaptation without forgetting. Furthermore, we propose a bi-directional guided modulation,

which is learned to automatically modulate the activations to reduce catastrophic forgetting. Extensive experimental results demonstrate that the proposed method outperforms the baseline and achieves the state-of-the-art results on CIFAR100, MiniImageNet, and CUB200 datasets.

****

The Majority Can Help the Minority: Context-Rich Minority Oversampling for Long-Tailed Classification

Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, Jin Young Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6887-6896

The problem of class imbalanced data is that the generalization performance of the classifier deteriorates due to the lack of data from minority classes. In this paper, we propose a novel minority over-sampling method to augment diversified minority samples by leveraging the rich context of the majority classes as background images. To diversify the minority samples, our key idea is to paste an image from a minority class onto rich-context images from a majority class, using them as background images. Our method is simple and can be easily combined with the existing long-tailed recognition methods. We empirically prove the effectiveness of the proposed oversampling method through extensive experiments and ablation studies. Without any architectural changes or complex algorithms, our method achieves state-of-the-art performance on various long-tailed classification benchmarks. Our code is made available at https://github.com/naver-ai/cmo.

****

Dense Depth Priors for Neural Radiance Fields From Sparse Input Views

Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12892-12901

Neural radiance fields (NeRF) encode a scene into a neural representation that enables photo-realistic rendering of novel views. However, a successful reconstruction from RGB images requires a large number of input views taken under static conditions - typically up to a few hundred images for room-size scenes. Our method aims to synthesize novel views of whole rooms from an order of magnitude fewer images. To this end, we leverage dense depth priors in order to constrain the NeRF optimization. First, we take advantage of the sparse depth data that is freely available from the structure from motion (SfM) preprocessing step used to estimate camera poses. Second, we use depth completion to convert these sparse points into dense depth maps and uncertainty estimates, which are used to guide NeRF optimization. Our method enables data-efficient novel view synthesis on challenging indoor scenes, using as few as 18 images for an entire scene.

****

EyePAD++: A Distillation-Based Approach for Joint Eye Authentication and Presentation Attack Detection Using Periocular Images

Prithviraj Dhar, Amit Kumar, Kirsten Kaplan, Khushi Gupta, Rakesh Ranjan, Rama Chellappa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20218-20227

A practical eye authentication (EA) system targeted for edge devices needs to perform authentication and be robust to presentation attacks, all while remaining compute and latency efficient. However, existing eye-based frameworks a) perform authentication and Presentation Attack Detection (PAD) independently and b) involve significant pre-processing steps to extract the iris region. Here, we introduce a joint framework for EA and PAD using periocular images. While a deep Multitask Learning (MTL) network can perform both the tasks, MTL suffers from the forgetting effect since the training datasets for EA and PAD are disjoint. To overcome this, we propose Eye Authentication with PAD (EyePAD), a distillation-based method that trains a single network for EA and PAD while reducing the effect of forgetting. To further improve the EA performance, we introduce a novel approach called EyePAD++ that includes training an MTL network on both EA and PAD data, while distilling the 'versatility' of the EyePAD network through an additional distillation step. Our proposed methods outperform the SOTA in PAD and obtain near-SOTA performance in eye-to-eye verification, without any pre-processing. We a

lso demonstrate the efficacy of EyePAD and EyePAD++ in user-to-user verification with PAD across network backbones and image quality.
****

IntentVizor: Towards Generic Query Guided Interactive Video Summarization
Guande Wu, Jianzhe Lin, Claudio T. Silva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10503-10512
The target of automatic video summarization is to create a short skim of the original long video while preserving the major content/events. There is a growing interest in the integration of user queries into video summarization or query-driven video summarization. This video summarization method predicts a concise synopsis of the original video based on the user query, which is commonly represented by the input text. However, two inherent problems exist in this query-driven way. First, the text query might not be enough to describe the exact and diverse needs of the user. Second, the user cannot edit once the summaries are produced, while we assume the needs of the user should be subtle and need to be adjusted interactively. To solve these two problems, we propose IntentVizor, an interactive video summarization framework guided by generic multi-modality queries. The input query that describes the user's needs are not limited to text but also the video snippets. We further represent these multi-modality finer-grained queries as user 'intent', which is interpretable, interactable, editable, and can better quantify the user's needs. In this paper, we use a set of the proposed intents to represent the user query and design a new interactive visual analytic interface. Users can interactively control and adjust these mixed-initiative intents to obtain a more satisfying summary through the interface. Also, to improve the summarization quality via video understanding, a novel Granularity-Scalable Ego-Graph Convolutional Networks (GSE-GCN) is proposed. We conduct our experiments on two benchmark datasets. Comparisons with the state-of-the-art methods verify the effectiveness of the proposed framework. Code and dataset are available at https://github.com/jnzs1836/intent-vizor.
****

Wnet: Audio-Guided Video Object Segmentation via Wavelet-Based Cross-Modal Denoising Networks
Wenwen Pan, Haonan Shi, Zhou Zhao, Jieming Zhu, Xiuqiang He, Zhigeng Pan, Lianli Gao, Jun Yu, Fei Wu, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1320-1331
Audio-Guided video semantic segmentation is a challenging problem in visual analysis and editing, which automatically separates foreground objects from background in a video sequence according to the referring audio expressions. However, the existing referring video semantic segmentation works mainly focus on the guidance of text-based referring expressions, due to the lack of modeling the semantic representation of audio-video interaction contents. In this paper, we consider the problem of audio-guided video semantic segmentation from the viewpoint of end-to-end denoised encoder-decoder network learning. We propose the walvelet-based encoder network to learn the crossmodal representations of the video contents with audio-form queries. Specifically, we adopt a multi-head cross-modal attention to explore the potential relations of video and query contents. A 2-dimension discrete wavelet transform is employed to decompose the audio-video features. We quantify the thresholds of high frequency coefficients to filter the noise and outliers. Then, a self attention-free decoder network is developed to generate the target masks with frequency domain transforms. Moreover, we maximize mutual information between the encoded features and multi-modal features after cross-modal attention to enhance the audio guidance. In addition, we construct the first large-scale audio-guided video semantic segmentation dataset. The extensive experiments show the effectiveness of our method.
****

Camera Pose Estimation Using Implicit Distortion Models
Linfei Pan, Marc Pollefeys, Viktor Larsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12819-12828
Low-dimensional parametric models are the de-facto standard in computer vision for intrinsic camera calibration. These models explicitly describe the mapping be

tween incoming viewing rays and image pixels. In this paper, we explore an alternative approach which implicitly models the lens distortion. The main idea is to replace the parametric model with a regularization term that ensures the latent distortion map varies smoothly throughout the image. The proposed model is effectively parameter-free and allows us to optimize the 6 degree-of-freedom camera pose without explicitly knowing the intrinsic calibration. We show that the method is applicable to a wide selection of cameras with varying distortion and in multiple applications, such as visual localization and structure-from-motion.
****

Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations
Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lu■i■, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6229-6238
A classical problem in computer vision is to infer a 3D scene representation from few images that can be used to render novel views at interactive rates. Previous work focuses on reconstructing pre-defined 3D representations, e.g. textured meshes, or implicit representations, e.g. radiance fields, and often requires input images with precise camera poses and long processing times for each novel scene. In this work, we propose the Scene Representation Transformer (SRT), a method which processes posed or unposed RGB images of a new area, infers a "set-latent scene representation", and synthesises novel views, all in a single feed-forward pass. To calculate the scene representation, we propose a generalization of the Vision Transformer to sets of images, enabling global information integration, and hence 3D reasoning. An efficient decoder transformer parameterizes the light field by attending into the scene representation to render novel views. Learning is supervised end-to-end by minimizing a novel-view reconstruction error. We show that this method outperforms recent baselines in terms of PSNR and speed on synthetic datasets, including a new dataset created for the paper. Further, we demonstrate that SRT scales to support interactive visualization and semantic segmentation of real-world outdoor environments using Street View imagery.
****

Shape-Invariant 3D Adversarial Point Clouds
Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15335-15344
Adversary and invisibility are two fundamental but conflict characters of adversarial perturbations. Previous adversarial attacks on 3D point cloud recognition have often been criticized for their noticeable point outliers, since they just involve an "implicit constrain" like global distance loss in the time-consuming optimization to limit the generated noise. While point cloud is a highly structured data format, it is hard to constrain its perturbation with a simple loss or metric properly. In this paper, we propose a novel Point-Cloud Sensitivity Map to boost both the efficiency and imperceptibility of point perturbations. This map reveals the vulnerability of point cloud recognition models when encountering shape-invariant adversarial noises. These noises are designed along the shape surface with an "explicit constrain" instead of extra distance loss. Specifically, we first apply a reversible coordinate transformation on each point of the point cloud input, to reduce one degree of point freedom and limit its movement on the tangent plane. Then we calculate the best attacking direction with the gradients of the transformed point cloud obtained on the white-box model. Finally we assign each point with a non-negative score to construct the sensitivity map, which benefits both white-box adversarial invisibility and black-box query-efficiency extended in our work. Extensive evaluations prove that our method can achieve the superior performance on various point cloud recognition models, with its satisfying adversarial imperceptibility and strong resistance to different point cloud defense settings. Our code is available at: https://github.com/shikiw/SI-Adv.
****

LAS-AT: Adversarial Training With Learnable Attack Strategy
Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, Xiaochun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13398-13408
Adversarial training (AT) is always formulated as a minimax problem, of which the performance depends on the inner optimization that involves the generation of adversarial examples (AEs). Most previous methods adopt Projected Gradient Descent (PGD) with manually specifying attack parameters for AE generation. A combination of the attack parameters can be referred to as an attack strategy. Several works have revealed that using a fixed attack strategy to generate AEs during the whole training phase limits the model robustness and propose to exploit different attack strategies at different training stages to improve robustness. But those multi-stage hand-crafted attack strategies need much domain expertise, and the robustness improvement is limited. In this paper, we propose a novel framework for adversarial training by introducing the concept of "learnable attack strategy", dubbed LAS-AT, which learns to automatically produce attack strategies to improve the model robustness. Our framework is composed of a target network that uses AEs for training to improve robustness, and a strategy network that produces attack strategies to control the AE generation. Experimental evaluations on three benchmark databases demonstrate the superiority of the proposed method, and the proposed method outperforms state-of-the-art adversarial training methods.
****
Bootstrapping ViTs: Towards Liberating Vision Transformers From Pre-Training
Haofei Zhang, Jiarui Duan, Mengqi Xue, Jie Song, Li Sun, Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8944-8953
Recently, vision Transformers (ViTs) are developing rapidly and starting to challenge the domination of convolutional neural networks (CNNs) in the realm of computer vision (CV). With the general-purpose Transformer architecture replacing the hard-coded inductive biases of convolution, ViTs have surpassed CNNs, especially in data-sufficient circumstances. However, ViTs are prone to over-fit on small datasets and thus rely on large-scale pre-training, which expends enormous time. In this paper, we strive to liberate ViTs from pre-training by introducing CNNs' inductive biases back to ViTs while preserving their network architectures for higher upper bound and setting up more suitable optimization objectives. To begin with, an agent CNN is designed based on the given ViT with inductive biases. Then a bootstrapping training algorithm is proposed to jointly optimize the agent and ViT with weight sharing, during which the ViT learns inductive biases from the intermediate features of the agent. Extensive experiments on CIFAR-10/100 and ImageNet-1k with limited training data have shown encouraging results that the inductive biases help ViTs converge significantly faster and outperform conventional CNNs with even fewer parameters. Our code is publicly available at https://github.com/zhfeing/Bootstrapping-ViTs-pytorch.
****
PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents
Brandon Smock, Rohith Pesala, Robin Abraham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4634-4642
Recently, significant progress has been made applying machine learning to the problem of table structure inference and extraction from unstructured documents. However, one of the greatest challenges remains the creation of datasets with complete, unambiguous ground truth at scale. To address this, we develop a new, more comprehensive dataset for table extraction, called PubTables-1M. PubTables-1M contains nearly one million tables from scientific articles, supports multiple input modalities, and contains detailed header and location information for table structures, making it useful for a wide variety of modeling approaches. It also addresses a significant source of ground truth inconsistency observed in prior datasets called oversegmentation, using a novel canonicalization procedure. We demonstrate that these improvements lead to a significant increase in training performance and a more reliable estimate of model performance at evaluation for table structure recognition. Further, we show that transformer-based object detect

ion models trained on PubTables-1M produce excellent results for all three tasks of detection, structure recognition, and functional analysis without the need for any special customization for these tasks. Data and code will be released at https://github.com/microsoft/table-transformer.

****

Styleformer: Transformer Based Generative Adversarial Networks With Style Vector
Jeeseung Park, Younggeun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8983-8992

We propose Styleformer, a generator that synthesizes image using style vectors based on the Transformer structure. In this paper, we effectively apply the modified Transformer structure (e.g., Increased multi-head attention and Pre-layer normalization) and attention style injection which is style modulation and demodulation method for self-attention operation. The new generator components have strengths in CNN's shortcomings, handling long-range dependency and understanding global structure of objects. We propose two methods to generate high-resolution images using Styleformer. First, we apply Linformer in the field of visual synthesis (Styleformer-L), enabling Styleformer to generate higher resolution images and result in improvements in terms of computation cost and performance. This is the first case using Linformer to image generation. Second, we combine Styleformer and StyleGAN2 (Styleformer-C) to generate high-resolution compositional scene efficiently, which Styleformer captures long-range-dependencies between components. With these adaptations, Styleformer achieves comparable performances to state-of-the-art in both single and multi-object datasets. Furthermore, groundbreaking results from style mixing and attention map visualization demonstrate the advantages and efficiency of our model.

****

Efficient Two-Stage Detection of Human-Object Interactions With a Novel Unary-Pairwise Transformer
Frederic Z. Zhang, Dylan Campbell, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20104-20112

Recent developments in transformer models for visual data have led to significant improvements in recognition and detection tasks. In particular, using learnable queries in place of region proposals has given rise to a new class of one-stage detection models, spearheaded by the Detection Transformer (DETR). Variations on this one-stage approach have since dominated human-object interaction (HOI) detection. However, the success of such one-stage HOI detectors can largely be attributed to the representation power of transformers. We discovered that when equipped with the same transformer, their two-stage counterparts can be more performant and memory-efficient, while taking a fraction of the time to train. In this work, we propose the Unary-Pairwise Transformer, a two-stage detector that exploits unary and pairwise representations for HOIs. We observe that the unary and pairwise parts of our transformer network specialise, with the former preferentially increasing the scores of positive examples and the latter decreasing the scores of negative examples. We evaluate our method on the HICO-DET and V-COCO datasets, and significantly outperform state-of-the-art approaches. At inference time, our model with ResNet50 approaches real-time performance on a single GPU.

****

ELSR: Efficient Line Segment Reconstruction With Planes and Points Guidance
Dong Wei, Yi Wan, Yongjun Zhang, Xinyi Liu, Bin Zhang, Xiqi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15807-15815

Three-dimensional (3D) line segments are helpful for scene reconstruction. Most of the existing 3D-line-segment-reconstruction algorithms deal with two views or dozens of small-size images; while in practice there are usually hundreds or thousands of large-size images. In this paper, we propose an efficient line segment reconstruction method called ELSR. ELSR exploits scene planes that are commonly seen in city scenes and sparse 3D points that can be acquired easily from the structure-from-motion (SfM) approach. For two views, ELSR efficiently finds the local scene plane to guide the line matching and exploits sparse 3D points to ac

celerate and constrain the matching. To reconstruct a 3D line segment with multi ple views, ELSR utilizes an efficient abstraction approach that selects represen tative 3D lines based on their spatial consistence. Our experiments demonstrated that ELSR had a higher accuracy and efficiency than the existing methods. Moreo ver, our results showed that ELSR could reconstruct 3D lines efficiently for lar ge and complex scenes that contain thousands of large-size images.
****

Meta-Attention for ViT-Backed Continual Learning
Mengqi Xue, Haofei Zhang, Jie Song, Mingli Song; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 150-159
Continual learning is a longstanding research topic due to its crucial role in t ackling continually arriving tasks. Up to now, the study of continual learning i n computer vision is mainly restricted to convolutional neural networks (CNNs). However, recently there is a tendency that the newly emerging vision transformer s (ViTs) are gradually dominating the field of computer vision, which leaves CNN -based continual learning lagging behind as they can suffer from severe performa nce degradation if straightforwardly applied to ViTs. In this paper, we study Vi T-backed continual learning to strive for higher performance riding on recent ad vances of ViTs. Inspired by mask-based continual learning methods in CNNs, where a mask is learned per task to adapt the pre-trained ViT to the new task, we pro pose MEta-ATtention (MEAT), i.e., attention to self-attention, to adapt a pre-tr ained ViT to new tasks without sacrificing performance on already learned tasks. Unlike prior mask-based methods like Piggyback, where all parameters are associ ated with corresponding masks, MEAT leverages the characteristics of ViTs and on ly masks a portion of its parameters. It renders MEAT more efficient and effecti ve with less overhead and higher accuracy. Extensive experiments demonstrate tha t MEAT exhibits significant superiority to its state-of-the-art CNN counterparts , with 4.0 6.0% absolute boosts in accuracy. Our code has been released at https ://github.com/zju-vipa/MEAT-TIL.
****

DST: Dynamic Substitute Training for Data-Free Black-Box Attack
Wenxuan Wang, Xuelin Qian, Yanwei Fu, Xiangyang Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14361-1 4370
With the wide applications of deep neural network models in various computer vis ion tasks, more and more works study the model vulnerability to adversarial exam ples. For data-free black box attack scenario, existing methods are inspired by the knowledge distillation, and thus usually train a substitute model to learn k nowledge from the target model using generated data as input. However, the subst itute model always has a static network structure, which limits the attack abili ty for various target models and tasks. In this paper, we propose a novel dynami c substitute training attack method to encourage substitute model to learn bette r and faster from the target model. Specifically, a dynamic substitute structure learning strategy is proposed to adaptively generate optimal substitute model s tructure via a dynamic gate according to different target models and tasks. More over, we introduce a task-driven graph-based structure information learning cons train to improve the quality of generated training data, and facilitate the subs titute model learning structural relationships from the target model multiple ou tputs. Extensive experiments have been conducted to verify the efficacy of the p roposed attack method, which can achieve better performance compared with the st ate-of-the-art competitors on several datasets. Project page: https://wxwangiris .github.io/DST
****

Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing
Thiemo Alldieck, Mihai Zanfir, Cristian Sminchisescu; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1506-1 515
We present PHORHUM, a novel, end-to-end trainable, deep neural network methodolo gy for photorealistic 3D human reconstruction given just a monocular RGB image. Our pixel-aligned method estimates detailed 3D geometry and, for the first time,

the unshaded surface color together with the scene illumination. Observing that 3D supervision alone is not sufficient for high fidelity color reconstruction, we introduce patch-based rendering losses that enable reliable color reconstruction on visible parts of the human, and detailed and plausible color estimation for the non-visible parts. Moreover, our method specifically addresses methodological and practical limitations of prior work in terms of representing geometry, albedo, and illumination effects, in an end-to-end model where factors can be effectively disentangled. In extensive experiments, we demonstrate the versatility and robustness of our approach. Our state-of-the-art results validate the method qualitatively and for different metrics, for both geometric and color reconstruction.

****

A Low-Cost & Real-Time Motion Capture System

Anargyros Chatzitofis, Georgios Albanis, Nikolaos Zioulis, Spyridon Thermos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21453-21458

Traditional marker-based motion capture requires excessive and specialized equipment, hindering accessibility and wider adoption. In this work, we demonstrate such a system but rely on a very sparse set of low-cost consumer-grade sensors. Our system exploits a data-driven backend to infer the captured subject's joint positions from noisy marker estimates in real-time. In addition to reduced costs and portability, its inherent denoising nature allows for quicker captures by alleviating the need for precise marker placement and post-processing, making it suitable for interactive virtual reality applications.

****

Unified Contrastive Learning in Image-Text-Label Space

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19163-19173

Visual recognition is recently learned via either supervised learning on human-annotated image-label data or language-image contrastive learning with webly-crawled image-text pairs. While supervised learning may result in a more discriminative representation, language-image pretraining shows unprecedented zero-shot recognition capability, largely due to the different data sources and learning objectives. In this work, we introduce a new formulation by combining the two data sources into a common image-text-label space. In this new space, we further propose a new learning method, called Unified Contrastive Learning (UniCL) with a single learning objective to seamlessly prompt the synergy between two types of data. Extensive experiments show that our UniCL is an effective way of learning semantically rich yet discriminative representations, universally for zero-shot, linear-probing, fully finetune and transfer learning scenarios. Particularly, it attains gains up to 9.2% and 14.5% in average on zero-shot recognition benchmarks over the language-image contrastive learning and supervised learning methods, respectively. In linear probing setting, it also boosts the performance over the two methods by 7.3% and 3.4%, respectively. Our further study indicates that UniCL is also a good learner on pure image-label data, rivaling the supervised learning methods across three image classification datasets and two types of vision backbone, ResNet and vision Transformer. ResNet and Swin Transformer. Code is available at: https://github.com/microsoft/UniCL.

****

Unifying Motion Deblurring and Frame Interpolation With Events

Xiang Zhang, Lei Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17765-17774

Slow shutter speed and long exposure time of frame-based cameras often cause visual blur and loss of inter-frame information, degenerating the overall quality of captured videos. To this end, we present a unified framework of event-based motion deblurring and frame interpolation for blurry video enhancement, where the extremely low latency of events is leveraged to alleviate motion blur and facilitate intermediate frame prediction. Specifically, the mapping relation between blurry frames and sharp latent images is first predicted by a learnable double in

tegral network, and a fusion network is then proposed to refine the coarse results via utilizing the information from consecutive blurry inputs and the concurrent events. By exploring the mutual constraints among blurry frames, latent images, and event streams, we further propose a self-supervised learning framework to enable network training with real-world blurry videos and events. Extensive experiments demonstrate that our method compares favorably against the state-of-the-art approaches and achieves remarkable performance on both synthetic and real-world datasets.

****

Generalizing Interactive Backpropagating Refinement for Dense Prediction Networks

Fanqing Lin, Brian Price, Tony Martinez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 773-782

As deep neural networks become the state-of-the-art approach in the field of computer vision for dense prediction tasks, many methods have been developed for automatic estimation of the target outputs given the visual inputs. Although the estimation accuracy of the proposed automatic methods continues to improve, interactive refinement is oftentimes necessary for further correction. Recently, feature backpropagating refinement scheme (f-BRS) has been proposed for the task of interactive segmentation, which enables efficient optimization of a small set of auxiliary variables inserted into the pretrained network to produce object segmentation that better aligns with user inputs. However, the proposed auxiliary variables only contain channel-wise scale and bias, limiting the optimization to global refinement only. In this work, in order to generalize backpropagating refinement for a wide range of dense prediction tasks, we introduce a set of G-BRS (Generalized Backpropagating Refinement Scheme) layers that enable both global and localized refinement for the following tasks: interactive segmentation, semantic segmentation, image matting and monocular depth estimation. Experiments on SBD, Cityscapes, Mapillary Vista, Composition-1k and NYU-Depth-V2 show that our method can successfully generalize and significantly improve performance of existing pretrained state-of-the-art models with only a few clicks.

****

Unsupervised Pre-Training for Temporal Action Localization Tasks

Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jue Wang, Yuexian Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14031-14041

Unsupervised video representation learning has made remarkable achievements in recent years. However, most existing methods are designed and optimized for video classification. These pre-trained models can be sub-optimal for temporal localization tasks due to the inherent discrepancy between video-level classification and clip-level localization. To bridge this gap, we make the first attempt to propose a self-supervised pretext task, coined as Pseudo Action Localization (PAL) to Unsupervisedly Pre-train feature encoders for Temporal Action Localization tasks (UP-TAL). Specifically, we first randomly select temporal regions, each of which contains multiple clips, from one video as pseudo actions and then paste them onto different temporal positions of the other two videos. The pretext task is to align the features of pasted pseudo action regions from two synthetic videos and maximize the agreement between them. Compared to the existing unsupervised video representation learning approaches, our PAL adapts better to downstream TAL tasks by introducing a temporal equivariant contrastive learning paradigm in a temporally dense and scale-aware manner. Extensive experiments show that PAL can utilize large-scale unlabeled video data to significantly boost the performance of existing TAL methods. Our codes and models will be made publicly available at https://github.com/zhang-can/UP-TAL.

****

Light Field Neural Rendering

Mohammed Suhail, Carlos Esteves, Leonid Sigal, Ameesh Makadia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8269-8279

Classical light field rendering for novel view synthesis can accurately reproduc

e view-dependent effects such as reflection, refraction, and translucency, but r equires a dense view sampling of the scene. Methods based on geometric reconstru ction need only sparse views, but cannot accurately model non-Lambertian effects . We introduce a model that combines the strengths and mitigates the limitations of these two directions. By operating on a four-dimensional representation of t he light field, our model learns to represent view-dependent effects accurately. By enforcing geometric constraints during training and inference, the scene geo metry is implicitly learned from a sparse set of views. Concretely, we introduce a two-stage transformer-based model that first aggregates features along epipol ar lines, then aggregates features along reference views to produce the color of a target ray. Our model outperforms the state-of-the-art on multiple forward-fa cing and 360deg datasets, with larger margins on scenes with severe view-depende nt variations.
****

Fast Point Transformer
Chunghyun Park, Yoonwoo Jeong, Minsu Cho, Jaesik Park; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16949 -16958
The recent success of neural networks enables a better interpretation of 3D poin t clouds, but processing a large-scale 3D scene remains a challenging problem. M ost current approaches divide a large-scale scene into small regions and combine the local predictions together. However, this scheme inevitably involves additi onal stages for pre- and post-processing and may also degrade the final output d ue to predictions in a local perspective. This paper introduces Fast Point Trans former that consists of a new lightweight self-attention layer. Our approach enc odes continuous 3D coordinates, and the voxel hashing-based architecture boosts computational efficiency. The proposed method is demonstrated with 3D semantic s egmentation and 3D detection. The accuracy of our approach is competitive to the best voxel based method, and our network achieves 129 times faster inference ti me than the state-of-the-art, Point Transformer, with a reasonable accuracy trad e-off in 3D semantic segmentation on S3DIS dataset.
****

Look Outside the Room: Synthesizing a Consistent Long-Term 3D Scene Video From a Single Image
Xuanchi Ren, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 3563-3573
Novel view synthesis from a single image has recently attracted a lot of attenti on, and it has been primarily advanced by 3D deep learning and rendering techniq ues. However, most work is still limited by synthesizing new views within relati vely small camera motions. In this paper, we propose a novel approach to synthes ize a consistent long-term video given a single scene image and a trajectory of large camera motions. Our approach utilizes an autoregressive Transformer to per form sequential modeling of multiple frames, which reasons the relations between multiple frames and the corresponding cameras to predict the next frame. To fac ilitate learning and ensure consistency among generated frames, we introduce a l ocality constraint based on the input cameras to guide self-attention among a la rge number of patches across space and time. Our method outperforms state-of-the -art view synthesis approaches by a large margin, especially when synthesizing l ong-term future in indoor 3D scenes. Project page at https://xrenaa.github.io/lo ok-outside-room/.
****

Unimodal-Concentrated Loss: Fully Adaptive Label Distribution Learning for Ordin al Regression
Qiang Li, Jingjing Wang, Zhaoliang Yao, Yachun Li, Pengju Yang, Jingwei Yan, Chu nmao Wang, Shiliang Pu; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 20513-20522
Learning from a label distribution has achieved promising results on ordinal reg ression tasks such as facial age and head pose estimation wherein, the concept o f adaptive label distribution learning (ALDL) has drawn lots of attention recent ly for its superiority in theory. However, compared with the methods assuming fi

xed form label distribution, ALDL methods have not achieved better performance. We argue that existing ALDL algorithms do not fully exploit the intrinsic properties of ordinal regression. In this paper, we emphatically summarize that learning an adaptive label distribution on ordinal regression tasks should follow three principles. First, the probability corresponding to the ground-truth should be the highest in label distribution. Second, the probabilities of neighboring labels should decrease with the increase of distance away from the ground-truth, i.e., the distribution is unimodal. Third, the label distribution should vary with samples changing, and even be distinct for different instances with the same label, due to the different levels of difficulty and ambiguity. Under the premise of these principles, we propose a novel loss function for fully adaptive label distribution learning, namely unimodal-concentrated loss. Specifically, the unimodal loss derived from the learning to rank strategy constrains the distribution to be unimodal. Furthermore, the estimation error and the variance of the predicted distribution for a specific sample are integrated into the proposed concentrated loss to make the predicted distribution maximize at the ground-truth and vary according to the predicting uncertainty. Extensive experimental results on typical ordinal regression tasks including age and head pose estimation, show the superiority of our proposed unimodal-concentrated loss compared with existing loss functions.

****

Augmented Geometric Distillation for Data-Free Incremental Person ReID

Yichen Lu, Mei Wang, Weihong Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7329-7338

Incremental learning (IL) remains an open issue for Person Re-identification (ReID), where a ReID system is expected to preserve preceding knowledge while learning incrementally. However, due to the strict privacy licenses and the open-set retrieval setting, it is intractable to adapt existing class IL methods to ReID. In this work, we propose an Augmented Geometric Distillation (AGD) framework to tackle these issues. First, a general data-free incremental framework with dreaming memory is constructed to avoid privacy disclosure. On this basis, we reveal a "noisy distillation" problem stemming from the noise in dreaming memory, and further propose to augment distillation in a pairwise and cross-wise pattern over different views of memory to mitigate it. Second, for the open-set retrieval property, we propose to maintain feature space structure during evolving via a novel geometric way and preserve relationships between exemplars when representations drift. Extensive experiments demonstrate the superiority of our AGD to baseline with a margin of 6.0% mAP / 7.9% R@1 and it could be generalized to class IL. Code is available.

****

Deep Stereo Image Compression via Bi-Directional Coding

Jianjun Lei, Xiangrui Liu, Bo Peng, Dengchao Jin, Wanqing Li, Jingxiao Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19669-19678

Existing learning-based stereo compression methods usually adopt a unidirectional approach to encoding one image independently and the other image conditioned upon the first. This paper proposes a novel bi-directional coding-based end-to-end stereo image compression network (BCSIC-Net). BCSIC-Net consists of a novel bi-directional contextual transform module which performs nonlinear transform conditioned upon the inter-view context in a latent space to reduce inter-view redundancy, and a bi-directional conditional entropy model that employs inter-view correspondence as a conditional prior to improve coding efficiency. Experimental results on the InStereo2K and KITTI datasets demonstrate that the proposed BCSIC-Net can effectively reduce the inter-view redundancy and outperforms state-of-the-art methods.

****

Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems Through Stochastic Contraction

Hyungjin Chung, Byeongsu Sim, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12413-12422

Diffusion models have recently attained significant interest within the community owing to their strong performance as generative models. Furthermore, its application to inverse problems have demonstrated state-of-the-art performance. Unfortunately, diffusion models have a critical downside - they are inherently slow to sample from, needing few thousand steps of iteration to generate images from pure Gaussian noise. In this work, we show that starting from Gaussian noise is unnecessary. Instead, starting from a single forward diffusion with better initialization significantly reduces the number of sampling steps in the reverse conditional diffusion. This phenomenon is formally explained by the contraction theory of the stochastic difference equations like our conditional diffusion strategy - the alternating applications of reverse diffusion followed by a non-expansive data consistency step. The new sampling strategy, dubbed Come-Closer-Diffuse-Faster (CCDF), also reveals a new insight on how the existing feed-forward neural network approaches for inverse problems can be synergistically combined with the diffusion models. Experimental results with super-resolution, image inpainting, and compressed sensing MRI demonstrate that our method can achieve state-of-the-art reconstruction performance at significantly reduced sampling steps.
****

Smooth-Swap: A Simple Enhancement for Face-Swapping With Smoothness
Jiseob Kim, Jihoon Lee, Byoung-Tak Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10779-10788
Face-swapping models have been drawing attention for their compelling generation quality, but their complex architectures and loss functions often require careful tuning for successful training. We propose a new face-swapping model called 'Smooth-Swap', which excludes complex handcrafted designs and allows fast and stable training. The main idea of Smooth-Swap is to build smooth identity embedding that can provide stable gradients for identity change. Unlike the one used in previous models trained for a purely discriminative task, the proposed embedding is trained with a supervised contrastive loss promoting a smoother space. With improved smoothness, Smooth-Swap suffices to be composed of a generic U-Net-based generator and three basic loss functions, a far simpler design compared with the previous models. Extensive experiments on face-swapping benchmarks (FFHQ, FaceForensics++) and face images in the wild show that our model is also quantitatively and qualitatively comparable or even superior to the existing methods.
****

Full-Range Virtual Try-On With Recurrent Tri-Level Transform
Han Yang, Xinrui Yu, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3460-3469
Virtual try-on aims to transfer a target clothing image onto a reference person. Though great progress has been achieved, the functioning zone of existing works is still limited to standard clothes (e.g., plain shirt without complex laces or ripped effect), while the vast complexity and variety of non-standard clothes (e.g., off-shoulder shirt, word-shoulder dress) are largely ignored. In this work, we propose a principled framework, Recurrent Tri-Level Transform (RT-VTON), that performs full-range virtual try-on on both standard and non-standard clothes. We have two key insights towards the framework design: 1) Semantics transfer requires a gradual feature transform on three different levels of clothing representations, namely clothes code, pose code and parsing code. 2) Geometry transfer requires a regularized image deformation between rigidity and flexibility. Firstly, we predict the semantics of the "after-try-on" person by recurrently refining the tri-level feature codes using local gated attention and non-local correspondence learning. Next, we design a semi-rigid deformation to align the clothing image and the predicted semantics, which preserves local warping similarity. Finally, a canonical try-on synthesizer fuses all the processed information to generate the clothed person image. Extensive experiments on conventional benchmarks along with user studies demonstrate that our framework achieves state-of-the-art performance both quantitatively and qualitatively. Notably, RT-VTON shows compelling results on a wide range of non-standard clothes.
****

Style Neophile: Constantly Seeking Novel Styles for Domain Generalization

Juwon Kang, Sohyun Lee, Namyup Kim, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7130-7140

This paper studies domain generalization via domain-invariant representation learning. Existing methods in this direction suppose that a domain can be characterized by styles of its images, and train a network using style-augmented data so that the network is not biased to particular style distributions. However, these methods are restricted to a finite set of styles since they obtain styles for augmentation from a fixed set of external images or by interpolating those of training data. To address this limitation and maximize the benefit of style augmentation, we propose a new method that synthesizes novel styles constantly during training. Our method manages multiple queues to store styles that have been observed so far, and synthesizes novel styles whose distribution is distinct from the distribution of styles in the queues. The style synthesis process is formulated as a monotone submodular optimization, thus can be conducted efficiently by a greedy algorithm. Extensive experiments on four public benchmarks demonstrate that the proposed method is capable of achieving state-of-the-art domain generalization performance.

****

High-Fidelity Human Avatars From a Single RGB Camera

Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, Kun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15904-15913

In this paper, we propose a coarse-to-fine framework to reconstruct a personalized high-fidelity human avatar from a monocular video. To deal with the misalignment problem caused by the changed poses and shapes in different frames, we design a dynamic surface network to recover pose-dependent surface deformations, which help to decouple the shape and texture of the person. To cope with the complexity of textures and generate photo-realistic results, we propose a reference-based neural rendering network and exploit a bottom-up sharpening-guided fine-tuning strategy to obtain detailed textures. Our framework also enables photo-realistic novel view/pose synthesis and shape editing applications. Experimental results on both the public dataset and our collected dataset demonstrate that our method outperforms the state-of-the-art methods. The code and dataset will be available at http://cic.tju.edu.cn/faculty/likun/projects/HF-Avatar.

****

ADAPT: Vision-Language Navigation With Modality-Aligned Action Prompts

Bingqian Lin, Yi Zhu, Zicong Chen, Xiwen Liang, Jianzhuang Liu, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15396-15406

Vision-Language Navigation (VLN) is a challenging task that requires an embodied agent to perform action-level modality alignment, i.e., make instruction-asked actions sequentially in complex visual environments. Most existing VLN agents learn the instruction-path data directly and cannot sufficiently explore action-level alignment knowledge inside the multi-modal inputs. In this paper, we propose modAlity-aligneD Action PrompTs (ADAPT), which provides the VLN agent with action prompts to enable the explicit learning of action-level modality alignment to pursue successful navigation. Specifically, an action prompt is defined as a modality-aligned pair of an image sub-prompt and a text sub-prompt, where the former is a single-view observation and the latter is a phrase like "walk past the chair". When starting navigation, the instruction-related action prompt set is retrieved from a pre-built action prompt base and passed through a prompt encoder to obtain the prompt feature. Then the prompt feature is concatenated with the original instruction feature and fed to a multi-layer transformer for action prediction. To collect high-quality action prompts into the prompt base, we use the Contrastive Language-Image Pretraining (CLIP) model which has powerful cross-modality alignment ability. A modality alignment loss and a sequential consistency loss are further introduced to enhance the alignment of the action prompt and enforce the agent to focus on the related prompt sequentially. Experimental results on both R2R and RxR show the superiority of ADAPT over state-of-the-art methods.

****
Multiview Transformers for Video Recognition
Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, Cordelia S chmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 3333-3343
Video understanding requires reasoning at multiple spatiotemporal resolutions -- from short fine-grained motions to events taking place over longer durations. A lthough transformer architectures have recently advanced the state-of-the-art, t hey have not explicitly modelled different spatiotemporal resolutions. To this e nd, we present Multiview Transformers for Video Recognition (MTV). Our model con sists of separate encoders to represent different views of the input video with lateral connections to fuse information across views. We present thorough ablati on studies of our model and show that MTV consistently performs better than sing le-view counterparts in terms of accuracy and computational cost across a range of model sizes. Furthermore, we achieve state-of-the-art results on six standard datasets, and improve even further with large-scale pretaining. Code and checkp oints are available at: https://github.com/google-research/scenic.
****
RIO: Rotation-Equivariance Supervised Learning of Robust Inertial Odometry
Xiya Cao, Caifa Zhou, Dandan Zeng, Yongliang Wang; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6614-6623
This paper introduces rotation-equivariance as a self-supervisor to train inerti al odometry models. We demonstrate that the self-supervised scheme provides a po werful supervisory signal at training phase as well as at inference stage. It re duces the reliance on massive amounts of labeled data for training a robust mode l and makes it possible to update the model using various unlabeled data. Furthe r, we propose adaptive Test-Time Training (TTT) based on uncertainty estimations in order to enhance the generalizability of the inertial odometry to various un seen data. We show in experiments that the Rotation-equivariance-supervised Iner tial Odometry (RIO) trained with 30% data achieves on par performance with a mod el trained with the whole database. Adaptive TTT improves models performance in all cases and makes more than 25% improvements under several scenarios.
****
How Good Is Aesthetic Ability of a Fashion Model?
Xingxing Zou, Kaicheng Pang, Wen Zhang, Waikeung Wong; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21200 -21209
We introduce A100 (Aesthetic 100) to assess the aesthetic ability of the fashion compatibility models. To date, it is the first work to address the AI model's a esthetic ability with detailed characterization based on the professional fashio n domain knowledge. A100 has several desirable characteristics: 1. Completeness. It covers all types of standards in the fashion aesthetic system through two te sts, namely LAT (Liberalism Aesthetic Test) and AAT (Academicism Aesthetic Test) ; 2. Reliability. It is training data agnostic and consistent with major indicat ors. It provides a fair and objective judgment for model comparison. 3. Explaina bility. Better than all previous indicators, the A100 further identifies essenti al characteristics of fashion aesthetics, thus showing the model's performance o n more fine-grained dimensions, such as Color, Balance, Material, etc. Experimen tal results prove the advance of the A100 in the aforementioned aspects. All dat a can be found at https://github.com/AemikaChow/AiDLab- fAshIon-Data.
****
Mining Multi-View Information: A Strong Self-Supervised Framework for Depth-Base d 3D Hand Pose and Mesh Estimation
Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, Jianxin Liao; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2022, pp. 20555-20565
In this work, we study the cross-view information fusion problem in the task of self-supervised 3D hand pose estimation from the depth image. Previous methods u sually adopt a hand-crafted rule to generate pseudo labels from multi-view estim ations in order to supervise the network training in each view. However, these m

ethods ignore the rich semantic information in each view and ignore the complex dependencies between different regions of different views. To solve these problems, we propose a cross-view fusion network to fully exploit and adaptively aggregate multi-view information. We encode diverse semantic information in each view into multiple compact nodes. Then, we introduce the graph convolution to model the complex dependencies between nodes and perform cross-view information interaction. Based on the cross-view fusion network, we propose a strong self-supervised framework for 3D hand pose and hand mesh estimation. Furthermore, we propose a pseudo multi-view training strategy to extend our framework to a more general scenario in which only single-view training data is used. Results on NYU dataset demonstrate that our method outperforms the previous self-supervised methods by 17.5% and 30.3% in multi-view and single-view scenarios. Meanwhile, our framework achieves comparable results to several strongly supervised methods.
****

Automated Progressive Learning for Efficient Training of Vision Transformers
Changlin Li, Bohan Zhuang, Guangrun Wang, Xiaodan Liang, Xiaojun Chang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12486-12496
Recent advances in vision Transformers (ViTs) have come with a voracious appetite for computing power, high-lighting the urgent need to develop efficient training methods for ViTs. Progressive learning, a training scheme where the model capacity grows progressively during training, has started showing its ability in efficient training. In this paper, we take a practical step towards efficient training of ViTs by customizing and automating progressive learning. First, we develop a strong manual baseline for progressive learning of ViTs, by introducing momentum growth (MoGrow) to bridge the gap brought by model growth. Then, we propose automated progressive learning (AutoProg), an efficient training scheme that aims to achieve lossless acceleration by automatically increasing the training overload on-the-fly; this is achieved by adaptively deciding whether, where and how much should the model grow during progressive learning. Specifically, we first relax the optimization of the growth schedule to sub-network architecture optimization problem, then propose one-shot estimation of the sub-network performance via an elastic supernet. The searching overhead is reduced to minimal by recycling the parameters of the supernet. Extensive experiments of efficient training on ImageNet with two representative ViT models, DeiT and VOLO, demonstrate that AutoProg can accelerate ViTs training by up to 85.1% with no performance drop.
****

BTS: A Bi-Lingual Benchmark for Text Segmentation in the Wild
Xixi Xu, Zhongang Qi, Jianqi Ma, Honglun Zhang, Ying Shan, Xiaohu Qie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19152-19162
As a prerequisite of many text-related tasks such as text erasing and text style transfer, text segmentation arouses more and more attention recently. Current researches mainly focus on only English characters and digits, while few work studies Chinese characters due to the lack of public large-scale and high-quality Chinese datasets, which limits the practical application scenarios of text segmentation. Different from English which has a limited alphabet of letters, Chinese has much more basic characters with complex structures, making the problem more difficult to deal with. To better analyze this problem, we propose the Bi-lingual Text Segmentation (BTS) dataset, a benchmark that covers various common Chinese scenes including 14,250 diverse and fine-annotated text images. BTS mainly focuses on Chinese characters, and also contains English words and digits. We also introduce Prior Guided Text Segmentation Network (PGTSNet), the first baseline to handle bi-lingual and complex-structured text segmentation. A plug-in text region highlighting module and a text perceptual discriminator are proposed in PGTSNet to supervise the model with text prior, and guide for more stable and finer text segmentation. A variation loss is also employed for suppressing background noise under complex scene. Extensive experiments are conducted not only to demonstrate the necessity and superiority of the proposed dataset BTS, but also to show the effectiveness of the proposed PGTSNet compared with a variety of state-of

-the-art text segmentation methods.
****

Learning Structured Gaussians To Approximate Deep Ensembles
Ivor J. A. Simpson, Sara Vicente, Neill D. F. Campbell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 366-374
This paper proposes using a sparse-structured multivariate Gaussian to provide a closed-form approximator for the output of probabilistic ensemble models used for dense image prediction tasks. This is achieved through a convolutional neural network that predicts the mean and covariance of the distribution, where the inverse covariance is parameterised by a sparsely structured Cholesky matrix. Similarly to distillation approaches, our single network is trained to maximise the probability of samples from pre-trained probabilistic models, in this work we use a fixed ensemble of networks. Once trained, our compact representation can be used to efficiently draw spatially correlated samples from the approximated output distribution. Importantly, this approach captures the uncertainty and structured correlations in the predictions explicitly in a formal distribution, rather than implicitly through sampling alone. This allows direct introspection of the model, enabling visualisation of the learned structure. Moreover, this formulation provides two further benefits: estimation of a sample probability, and the introduction of arbitrary spatial conditioning at test time. We demonstrate the merits of our approach on monocular depth estimation and show that the advantages of our approach are obtained with comparable quantitative performance.
****

Adaptive Trajectory Prediction via Transferable GNN
Yi Xu, Lichen Wang, Yizhou Wang, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6520-6531
Pedestrian trajectory prediction is an essential component in a wide range of AI applications such as autonomous driving and robotics. Existing methods usually assume the training and testing motions follow the same pattern while ignoring the potential distribution differences (e.g., shopping mall and street). This issue results in inevitable performance decrease. To address this issue, we propose a novel Transferable Graph Neural Network (T-GNN) framework, which jointly conducts trajectory prediction as well as domain alignment in a unified framework. Specifically, a domain-invariant GNN is proposed to explore the structural motion knowledge where the domain-specific knowledge is reduced. Moreover, an attention-based adaptive knowledge learning module is further proposed to explore fine-grained individual-level feature representations for knowledge transfer. By this way, disparities across different trajectory domains will be better alleviated. More challenging while practical trajectory prediction experiments are designed, and the experimental results verify the superior performance of our proposed model. To the best of our knowledge, our work is the pioneer which fills the gap in benchmarks and techniques for practical pedestrian trajectory prediction across different domains.
****

Total Variation Optimization Layers for Computer Vision
Raymond A. Yeh, Yuan-Ting Hu, Zhongzheng Ren, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 711-721
Optimization within a layer of a deep-net has emerged as a new direction for deep-net layer design. However, there are two main challenges when applying these layers to computer vision tasks: (a) which optimization problem within a layer is useful?; (b) how to ensure that computation within a layer remains efficient? To study question (a), in this work, we propose total variation (TV) minimization as a layer for computer vision. Motivated by the success of total variation in image processing, we hypothesize that TV as a layer provides useful inductive bias for deep-nets too. We study this hypothesis on five computer vision tasks: image classification, weakly-supervised object localization, edge-preserving smoothing, edge detection, and image denoising, improving over existing baselines. To achieve these results, we had to address question (b): we developed a GPU-based

projected-Newton method which is 37x faster than existing solutions.
****

Defensive Patches for Robust Recognition in the Physical World
Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2456-2465
To operate in real-world high-stakes environments, deep learning systems have to endure noises that have been continuously thwarting their robustness. Data-end defense, which improves robustness by operations on input data instead of modifying models, has attracted intensive attention due to its high feasibility in practice. However, previous data-end defenses show low generalization against diverse noises and weak transferability across multiple models. Motivated by the fact that robust recognition depends on both local and global features, we propose a defensive patch generation framework to address these problems by helping models better exploit these features. For the generalization against diverse noises, we inject class-specific identifiable patterns into a confined local patch prior, so that defensive patches could preserve more recognizable features towards specific classes, leading models for better recognition under noises. For the transferability across multiple models, we guide the defensive patches to capture more global feature correlations within a class, so that they could activate model-shared global perceptions and transfer better among models. Our defensive patches show great potentials to improve model robustness in practice by simply sticking them around target objects. Extensive experiments show that we outperform others by large margins (improve 20+% accuracy for both adversarial and corruption robustness on average in the digital and physical world).
****

Single-Stage Is Enough: Multi-Person Absolute 3D Pose Estimation
Lei Jin, Chenyang Xu, Xiaojuan Wang, Yabo Xiao, Yandong Guo, Xuecheng Nie, Jian Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13086-13095
The existing multi-person absolute 3D pose estimation methods are mainly based on two-stage paradigm, i.e., top-down or bottom-up, leading to redundant pipelines with high computation cost. We argue that it is more desirable to simplify such two-stage paradigm to a single-stage one to promote both efficiency and performance. To this end, we present an efficient single-stage solution, Decoupled Regression Model (DRM), with three distinct novelties. First, DRM introduces a new decoupled representation for 3D pose, which expresses the 2D pose in image plane and depth information of each 3D human instance via 2D center point (center of visible keypoints) and root point (denoted as pelvis), respectively. Second, to learn better feature representation for the human depth regression, DRM introduces a 2D Pose-guided Depth Query Module (PDQM) to extract the features in 2D pose regression branch, enabling the depth regression branch to perceive the scale information of instances. Third, DRM leverages a Decoupled Absolute Pose Loss (DAPL) to facilitate the absolute root depth and root-relative depth estimation, thus improving the accuracy of absolute 3D pose. Comprehensive experiments on challenging benchmarks including MuPoTS-3D and Panoptic clearly verify the superiority of our framework, which outperforms the state-of-the-art bottom-up absolute 3D pose estimation methods.
****

Deformation and Correspondence Aware Unsupervised Synthetic-to-Real Scene Flow Estimation for Point Clouds
Zhao Jin, Yinjie Lei, Naveed Akhtar, Haifeng Li, Munawar Hayat; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7233-7243
Point cloud scene flow estimation is of practical importance for dynamic scene navigation in autonomous driving. Since scene flow labels are hard to obtain, current methods train their models on synthetic data and transfer them to real scenes. However, large disparities between existing synthetic datasets and real scenes lead to poor model transfer. We make two major contributions to address that. First, we develop a point cloud collector and scene flow annotator for GTA-V en

gine to automatically obtain diverse realistic training samples without human in tervention. With that, we develop a large-scale synthetic scene flow dataset GTA -SF. Second, we propose a mean-teacher-based domain adaptation framework that le verages self-generated pseudo-labels of the target domain. It also explicitly in corporates shape deformation regularization and surface correspondence refinemen t to address distortions and misalignments in domain transfer. Through extensive experiments, we show that our GTA-SF dataset leads to a consistent boost in mod el generalization to three real datasets (i.e., Waymo, Lyft and KITTI) as compar ed to the most widely used FT3D dataset. Moreover, our framework achieves superi or adaptation performance on six source-target dataset pairs, remarkably closing the average domain gap by 60%. Data and codes are available at https://github.c om/leolyj/DCA-SRSFE

****

Learn From Others and Be Yourself in Heterogeneous Federated Learning
Wenke Huang, Mang Ye, Bo Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10143-10153

Federated learning has emerged as an important distributed learning paradigm, wh ich normally involves collaborative updating with others and local updating on p rivate data. However, heterogeneity problem and catastrophic forgetting bring di stinctive challenges. First, due to non-i.i.d (identically and independently dis tributed) data and heterogeneous architectures, models suffer performance degrad ation on other domains and communication barrier with participants models. Secon d, in local updating, model is separately optimized on private data, which is pr one to overfit current data distribution and forgets previously acquired knowled ge, resulting in catastrophic forgetting. In this work, we propose FCCL (Federat ed Cross-Correlation and Continual Learning). For heterogeneity problem, FCCL le verages unlabeled public data for communication and construct cross-correlation matrix to learn a generalizable representation under domain shift. Meanwhile, fo r catastrophic forgetting, FCCL utilizes knowledge distillation in local updatin g, providing inter and intra domain information without leaking privacy. Empiric al results on various image classification tasks demonstrate the effectiveness o f our method and the efficiency of modules.

****

Sequential Voting With Relational Box Fields for Active Object Detection
Qichen Fu, Xingyu Liu, Kris Kitani; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2022, pp. 2374-2383

A key component of understanding hand-object interactions is the ability to iden tify the active object -- the object that is being manipulated by the human hand . In order to accurately localize the active object, any method must reason usin g information encoded by each image pixel, such as whether it belongs to the han d, the object, or the background. To leverage each pixel as evidence to determin e the bounding box of the active object, we propose a pixel-wise voting function . Our pixel-wise voting function takes an initial bounding box as input and prod uces an improved bounding box of the active object as output. The voting functio n is designed so that each pixel inside of the input bounding box votes for an i mproved bounding box, and the box with the majority vote is selected as the outp ut. We call the collection of bounding boxes generated inside of the voting func tion, the Relational Box Field, as it characterizes a field of bounding boxes de fined in relationship to the current bounding box. While our voting function is able to improve the bounding box of the active object, one round of voting is ty pically not enough to accurately localize the active object. Therefore, we repea tedly apply the voting function to sequentially improve the location of the boun ding box. However, since it is known that repeatedly applying a one-step predict or (i.e., auto-regressive processing with our voting function) can cause a data distribution shift, we mitigate this issue using reinforcement learning (RL). We adopt standard RL to learn the voting function parameters and show that it prov ides a meaningful improvement over a standard supervised learning approach. We p erform experiments on two large-scale datasets: 100DOH and MECCANO, improving AP 50 performance by 8% and 30%, respectively, over the state of the art. The proje ct page with code and visualizations can be found at https://fuqichen1998.github

.io/SequentialVotingDet/.
****

Semantic-Aware Auto-Encoders for Self-Supervised Representation Learning
Guangrun Wang, Yansong Tang, Liang Lin, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9664-9675
The resurgence of unsupervised learning can be attributed to the remarkable progress of self-supervised learning, which includes generative (G) and discriminative (D) models. In computer vision, the mainstream self-supervised learning algorithms are D models. However, designing a D model could be over-complicated; also, some studies hinted that a D model might not be as general and interpretable as a G model. In this paper, we switch from D models to G models using the classical auto-encoder (AE). Note that a vanilla G model was far less efficient than a D model in self-supervised computer vision tasks, as it wastes model capability on overfitting semantic-agnostic high-frequency details. Inspired by perceptual learning that could use cross-view learning to perceive concepts and semantics, we propose a novel AE that could learn semantic-aware representation via cross-view image reconstruction. We use one view of an image as the input and another view of the same image as the reconstruction target. This kind of AE has rarely been studied before, and the optimization is very difficult. To enhance learning ability and find a feasible solution, we propose a semantic aligner that uses geometric transformation knowledge to align the hidden code of AE to help optimization. These techniques significantly improve the representation learning ability of AE and make self-supervised learning with G models possible. Extensive experiments on many large-scale benchmarks (e.g., ImageNet, COCO 2017, and SYSU-30k) demonstrate the effectiveness of our methods. Code is available at https://github.com/wanggrun/Semantic-Aware-AE.
****

Learning Transferable Human-Object Interaction Detector With Natural Language Supervision
Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 939-948
It is difficult to construct a data collection including all possible combinations of human actions and interacting objects due to the combinatorial nature of human-object interactions (HOI). In this work, we aim to develop a transferable HOI detector for unseen interactions. Existing HOI detectors often treat interactions as discrete labels and learn a classifier according to a predetermined category space. This is inherently inapt for detecting unseen interactions which are out of the predefined categories. Conversely, we treat independent HOI labels as the natural language supervision of interactions and embed them into a joint visual-and-text space to capture their correlations. More specifically, we propose a new HOI visual encoder to detect the interacting humans and objects, and map them to a joint feature space to perform interaction recognition. Our visual encoder is instantiated as a Vision Transformer with new learnable HOI tokens and a sequence parser to generate unique HOI predictions. It distills and leverages the transferable knowledge from the pretrained CLIP model to perform the zero-shot interaction detection. Experiments on two datasets, SWIG-HOI and HICO-DET, validate that our proposed method can achieve a notable mAP improvement on detecting both seen and unseen HOIs.
****

Fourier Document Restoration for Robust Document Dewarping and Recognition
Chuhui Xue, Zichen Tian, Fangneng Zhan, Shijian Lu, Song Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4573-4582
State-of-the-art document dewarping techniques learn to predict 3-dimensional information of documents which are prone to errors while dealing with documents with irregular distortions or large variations in depth. This paper presents FDRNet, a Fourier Document Restoration Network that can restore documents with different distortions and improve document recognition in a reliable and simpler manne

r. FDRNet focuses on high-frequency components in the Fourier space that capture most structural information but are largely free of degradation in appearance. It dewarps documents by a flexible Thin-Plate Spline transformation which can handle various deformations effectively without requiring deformation annotations in training. These features allow FDRNet to learn from a small amount of simply labeled training images, and the learned model can dewarp documents with complex geometric distortion and recognize the restored texts accurately. To facilitate document restoration research, we create a benchmark dataset consisting of over one thousand camera documents with different types of geometric and photometric distortion. Extensive experiments show that FDRNet outperforms the state-of-the-art by large margins on both dewarping and text recognition tasks. In addition, FDRNet requires a small amount of simply labeled training data and is easy to deploy.

****

Consistency Learning via Decoding Path Augmentation for Transformers in Human Object Interaction Detection

Jihwan Park, SeungJun Lee, Hwan Heo, Hyeong Kyu Choi, Hyunwoo J. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1019-1028

Human-Object Interaction detection is a holistic visual recognition task that entails object detection as well as interaction classification. Previous works of HOI detection has been addressed by the various compositions of subset predictions, e.g., Image -> HO -> I, Image -> HI -> O. Recently, transformer based architecture for HOI has emerged, which directly predicts the HOI triplets in an end-to-end fashion (Image -> HOI). Motivated by various inference paths for HOI detection, we propose cross-path consistency learning (CPC), which is a novel end-to-end learning strategy to improve HOI detection for transformers by leveraging augmented decoding paths. CPC learning enforces all the possible predictions from permuted inference sequences to be consistent. This simple scheme makes the model learn consistent representations, thereby improving generalization without increasing model capacity. Our experiments demonstrate the effectiveness of our method, and we achieved significant improvement on V-COCO and HICO-DET compared to the baseline models. Our code is available at https://github.com/mlvlab/CPChoi.

****

Consistent Explanations by Contrastive Learning

Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, Hamed Pirsiavash; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10213-10222

Post-hoc explanation methods, e.g., Grad-CAM, enable humans to inspect the spatial regions responsible for a particular network decision. However, it is shown that such explanations are not always consistent with human priors, such as consistency across image transformations. Given an interpretation algorithm, e.g., Grad-CAM, we introduce a novel training method to train the model to produce more consistent explanations. Since obtaining the ground truth for a desired model interpretation is not a well-defined task, we adopt ideas from contrastive self-supervised learning, and apply them to the interpretations of the model rather than its embeddings. We show that our method, Contrastive Grad-CAM Consistency (CGC), results in Grad-CAM interpretation heatmaps that are more consistent with human annotations while still achieving comparable classification accuracy. Moreover, our method acts as a regularizer and improves the accuracy on limited-data, fine-grained classification settings. In addition, because our method does not rely on annotations, it allows for the incorporation of unlabeled data into training, which enables better generalization of the model. The code is available here: https://github.com/UCDvision/CGC

****

Text2Pos: Text-to-Point-Cloud Cross-Modal Localization

Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, Laura Leal-Taixé; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6687-6696

Natural language-based communication with mobile devices and home appliances is

becoming increasingly popular and has the potential to become natural for communicating with mobile robots in the future. Towards this goal, we investigate cross-modal text-to-point-cloud localization that will allow us to specify, for example, a vehicle pick-up or goods delivery location. In particular, we propose Text2Pos, a cross-modal localization module that learns to align textual descriptions with localization cues in a coarse- to-fine manner. Given a point cloud of the environment, Text2Pos locates a position that is specified via a natural language-based description of the immediate surroundings. To train Text2Pos and study its performance, we construct KITTI360Pose, the first dataset for this task based on the recently introduced KITTI360 dataset. Our experiments show that we can localize 65% of textual queries within 15m distance to query locations for top-10 retrieved locations. This is a starting point that we hope will spark future developments towards language-based navigation.

****

MulT: An End-to-End Multitask Learning Transformer

Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12031-12041

We propose an end-to-end Multitask Learning Transformer framework, named MulT, to simultaneously learn multiple high-level vision tasks, including depth estimation, semantic segmentation, reshading, surface normal estimation, 2D keypoint detection, and edge detection. Based on the Swin transformer model, our framework encodes the input image into a shared representation and makes predictions for each vision task using task-specific transformer-based decoder heads. At the heart of our approach is a shared attention mechanism modeling the dependencies across the tasks. We evaluate our model on several multitask benchmarks, showing that our MulT framework outperforms both the state-of-the art multitask convolutional neural network models and all the respective single task transformer models. Our experiments further highlight the benefits of sharing attention across all the tasks, and demonstrate that our MulT model is robust and generalizes well to new domains. We will make our code and models publicly available upon publication.

****

Hierarchical Modular Network for Video Captioning

Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17939-17948

Video captioning aims to generate natural language descriptions according to the content, where representation learning plays a crucial role. Existing methods are mainly developed within the supervised learning framework via word-by-word comparison of the generated caption against the ground-truth text without fully exploiting linguistic semantics. In this work, we propose a hierarchical modular network to bridge video representations and linguistic semantics from three levels before generating captions. In particular, the hierarchy is composed of: (I) Entity level, which highlights objects that are most likely to be mentioned in captions. (II) Predicate level, which learns the actions conditioned on highlighted objects and is supervised by the predicate in captions. (III) Sentence level, which learns the global semantic representation and is supervised by the whole caption. Each level is implemented by one module. Extensive experimental results show that the proposed method performs favorably against the state-of-the-art models on the two widely-used benchmarks: MSVD 104.0% and MSR-VTT 51.5% in CIDEr score. Code will be made available at https://github.com/MarcusNerva/HMN.

****

Learning With Neighbor Consistency for Noisy Labels

Ahmet Iscen, Jack Valmadre, Anurag Arnab, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4672-4681

Recent advances in deep learning have relied on large, labelled datasets to train high-capacity models. However, collecting large datasets in a time- and cost-efficient manner often results in label noise. We present a method for learning f

rom noisy labels that leverages similarities between training examples in featur e space, encouraging the prediction of each example to be similar to its nearest neighbours. Compared to training algorithms that use multiple models or distinc t stages, our approach takes the form of a simple, additional regularization ter m. It can be interpreted as an inductive version of the classical, transductive label propagation algorithm. We thoroughly evaluate our method on datasets evalu ating both synthetic (CIFAR-10, CIFAR-100) and realistic (mini-WebVision, WebVis ion, Clothing1M, mini-ImageNet-Red) noise, and achieve competitive or state-of-t he-art accuracies across all of them.

****

Depth Estimation by Combining Binocular Stereo and Monocular Structured-Light

Yuhua Xu, Xiaoli Yang, Yushan Yu, Wei Jia, Zhaobi Chu, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 1746-1755

It is well known that the passive stereo system cannot adapt well to weak textur e objects, e.g., white walls. However, these weak texture targets are very commo n in indoor environments. In this paper, we present a novel stereo system, which consists of two cameras (an RGB camera and an IR camera) and an IR speckle proj ector. The RGB camera is used both for depth estimation and texture acquisition. The IR camera and the speckle projector can form a monocular structured-light (MSL) subsystem, while the two cameras can form a binocular stereo subsystem. The depth map generated by the MSL subsystem can provide external guidance for the stereo matching networks, which can improve the matching accuracy significantly. In order to verify the effectiveness of the proposed system, we built a prototy pe and collected a test dataset in indoor scenes. The evaluation results show th at the Bad 2.0 error of the proposed system is 28.2% of the passive stereo syste m when the network RAFT is used. The dataset and trained models are available at https://github.com/YuhuaXu/MonoStereoFusion.

****

Salient-to-Broad Transition for Video Person Re-Identification

Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, Xilin Chen; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7339-7348

Due to the limited utilization of temporal relations in video re-id, the frame-l evel attention regions of mainstream methods are partial and highly similar. To address this problem, we propose a Salient-to-Broad Module (SBM) to enlarge the attention regions gradually. Specifically, in SBM, while the previous frames hav e focused on the most salient regions, the later frames tend to focus on broader regions. In this way, the additional information in broad regions can supplemen t salient regions, incurring more powerful video-level representations. To furth er improve SBM, an Integration-and-Distribution Module (IDM) is introduced to en hance frame-level representations. IDM first integrates features from the entire feature space and then distributes the integrated features to each spatial loca tion. SBM and IDM are mutually beneficial since they enhance the representations from video-level and framelevel, respectively. Extensive experiments on four pr evalent benchmarks demonstrate the effectiveness and superiority of our method. The source code is available at https://github.com/baist/SINet.

****

Object-Region Video Transformers

Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, Amir Globerson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3148-3159

Recently, video transformers have shown great success in video understanding, ex ceeding CNN performance; yet existing video transformer models do not explicitly model objects, although objects can be essential for recognizing actions. In th is work, we present Object-Region Video Transformers (ORViT), an object-centric approach that extends video transformer layers with a block that directly incorp orates object representations. The key idea is to fuse object-centric representa tions starting from early layers and propagate them into the transformer-layers, thus affecting the spatio-temporal representations throughout the network. Our

ORViT block consists of two object-level streams: appearance and dynamics. In th e appearance stream, an "Object-Region Attention" module applies self-attention over the patches and object regions. In this way, visual object regions interact with uniform patch tokens and enrich them with contextualized object informatio n. We further model object dynamics via a separate "Object-Dynamics Module", whi ch captures trajectory interactions, and show how to integrate the two streams. We evaluate our model on four tasks and five datasets: compositional and few-sho t action recognition on SomethingElse, spatio-temporal action detection on AVA, and standard action recognition on Something-Something V2, Diving48 and Epic-Kit chen100. We show strong performance improvement across all tasks and datasets co nsidered, demonstrating the value of a model that incorporates object representa tions into a transformer architecture.
****

DeeCap: Dynamic Early Exiting for Efficient Image Captioning
Zhengcong Fei, Xu Yan, Shuhui Wang, Qi Tian; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12216-12226
Both accuracy and efficiency are crucial for image captioning in real-world scen arios. Although Transformer-based models have gained significant improved captio ning performance, their computational cost is very high. A feasible way to reduc e the time complexity is to exit the prediction early in internal decoding layer s without passing the entire model. However, it is not straightforward to devise early exiting into image captioning due to the following issues. On one hand, t he representation in shallow layers lacks high-level semantic and sufficient cro ss-modal fusion information for accurate prediction. On the other hand, the exit ing decisions made by internal classifiers are unreliable sometimes. To solve th ese issues, we propose DeeCap framework for efficient image captioning, which dy namically selects proper-sized decoding layers from a global perspective to exit early. The key to successful early exiting lies in the specially designed imita tion learning mechanism, which predicts the deep layer activation with shallow l ayer features. By deliberately merging the imitation learning into the whole ima ge captioning architecture, the imitated deep layer representation can mitigate the loss brought by the missing of actual deep layers when early exiting is unde rtaken, resulting in significant reduction in calculation cost with small sacrif ice of accuracy. Experiments on the MS COCO and Flickr30k datasets demonstrate t he DeeCap can achieve competitive performances with 4 speed-up. Code is availabl e at: https://github.com/feizc/DeeCap.
****

AME: Attention and Memory Enhancement in Hyper-Parameter Optimization
Nuo Xu, Jianlong Chang, Xing Nie, Chunlei Huo, Shiming Xiang, Chunhong Pan; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 480-489
Training Deep Neural Networks (DNNs) is inherently subject to sensitive hyper-pa rameters and untimely feedbacks of performance evaluation. To solve these two di fficulties, an efficient parallel hyper-parameter optimization model is proposed under the framework of Deep Reinforcement Learning (DRL). Technically, we devel op Attention and Memory Enhancement (AME), that includes multi-head attention an d memory mechanism to enhance the ability to capture both the short-term and lon g-term relationships between different hyper-parameter configurations, yielding an attentive sampling mechanism for searching high-performance configurations em bedded into a huge search space. During the optimization of transformer-structur ed configuration searcher, a conceptually intuitive yet powerful strategy is app lied to solve the problem of insufficient number of samples due to the untimely feedback. Experiments on three visual tasks, including image classification, obj ect detection, semantic segmentation, demonstrate the effectiveness of AME.
****

Alignment-Uniformity Aware Representation Learning for Zero-Shot Video Classific ation
Shi Pu, Kaili Zhao, Mao Zheng; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 19968-19977
Most methods tackle zero-shot video classification by aligning visual-semantic r

epresentations within seen classes, which limits generalization to unseen classes. To enhance model generalizability, this paper presents an end-to-end framework that preserves alignment and uniformity properties for representations on both seen and unseen classes. Specifically, we formulate a supervised contrastive loss to simultaneously align visual-semantic features (i.e., alignment) and encourage the learned features to distribute uniformly (i.e., uniformity). Unlike existing methods that only consider the alignment, we propose uniformity to preserve maximal-info of existing features, which improves the probability that unobserved features fall around observed data. Further, we synthesize features of unseen classes by proposing a class generator that interpolates and extrapolates the features of seen classes. Besides, we introduce two metrics, closeness and dispersion, to quantify the two properties and serve as new measurements of model generalizability. Experiments show that our method significantly outperforms SoTA by relative improvements of 28.1% on UCF101 and 27.0% on HMDB51. Code is available.

****

RepMLPNet: Hierarchical Vision MLP With Re-Parameterized Locality

Xiaohan Ding, Honghao Chen, Xiangyu Zhang, Jungong Han, Guiguang Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 578-587

Compared to convolutional layers, fully-connected (FC) layers are better at modeling the long-range dependencies but worse at capturing the local patterns, hence usually less favored for image recognition. In this paper, we propose a methodology, Locality Injection, to incorporate local priors into an FC layer via merging the trained parameters of a parallel conv kernel into the FC kernel. Locality Injection can be viewed as a novel Structural Re-parameterization method since it equivalently converts the structures via transforming the parameters. Based on that, we propose a multi-layer-perceptron (MLP) block named RepMLP Block, which uses three FC layers to extract features, and a novel architecture named RepMLPNet. The hierarchical design distinguishes RepMLPNet from the other concurrently proposed vision MLPs. As it produces feature maps of different levels, it qualifies as a backbone model for downstream tasks like semantic segmentation. Our results reveal that 1) Locality Injection is a general methodology for MLP models; 2) RepMLPNet has favorable accuracy-efficiency trade-off compared to the other MLPs; 3) RepMLPNet is the first MLP that seamlessly transfer to Cityscapes semantic segmentation. The code and models are available at https://github.com/DingXiaoH/RepMLP.

****

DR.VIC: Decomposition and Reasoning for Video Individual Counting

Tao Han, Lei Bai, Junyu Gao, Qi Wang, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3083-3092

Pedestrian counting is a fundamental tool for understanding pedestrian patterns and crowd flow analysis. Existing works (e.g., image-level pedestrian counting, crossline crowd counting et al.) either only focus on the image-level counting or are constrained to the manual annotation of lines. In this work, we propose to conduct the pedestrian counting from a new perspective - Video Individual Counting (VIC), which counts the total number of individual pedestrians in the given video (a person is only counted once). Instead of relying on the Multiple Object Tracking (MOT) techniques, we propose to solve the problem by decomposing all pedestrians into the initial pedestrians who existed in the first frame and the new pedestrians with separate identities in each following frame. Then, an end-to-end Decomposition and Reasoning Network (DRNet) is designed to predict the initial pedestrian count with the density estimation method and reason the new pedestrian's count of each frame with the differentiable optimal transport. Extensive experiments are conducted on two datasets with congested pedestrians and diverse scenes, demonstrating the effectiveness of our method over baselines with great superiority in counting the individual pedestrians. Code: https://github.com/taohan10200/DRNet.

****

LiDARCap: Long-Range Marker-Less 3D Human Motion Capture With LiDAR Point Clouds
Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20502-20512
Existing motion capture datasets are largely short-range and cannot yet fit the need of long-range applications. We propose LiDARHuman26M, a new human motion capture dataset captured by LiDAR at a much longer range to overcome this limitation. Our dataset also includes the ground truth human motions acquired by the IMU system and the synchronous RGB images. We further present a strong baseline method, LiDARCap, for LiDAR point cloud human motion capture. Specifically, we first utilize PointNet++ to encode features of points and then employ the inverse kinematics solver and SMPL optimizer to regress the pose through aggregating the temporally encoded features hierarchically. Quantitative and qualitative experiments show that our method outperforms the techniques based only on RGB images. Ablation experiments demonstrate that our dataset is challenging and worthy of further research. Finally, the experiments on the KITTI Dataset and the Waymo Open Dataset show that our method can be generalized to different LiDAR sensor settings.
****
GeoEngine: A Platform for Production-Ready Geospatial Research
Sagar Verma, Siddharth Gupta, Hal Shin, Akash Panigrahi, Shubham Goswami, Shweta Pardeshi, Natanael Exe, Ujwal Dutta, Tanka Raj Joshi, Nitin Bhojwani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21416-21424
Geospatial machine learning has seen tremendous academic advancement, but its practical application has been constrained by difficulties with operationalizing performant and reliable solutions. Sourcing satellite imagery in real-world settings, handling terabytes of training data, and managing machine learning artifacts are a few of the challenges that have severely limited downstream innovation. In this paper we introduce the GeoEngine platform for reproducible and production-ready geospatial machine learning research. GeoEngine removes key technical hurdles to adopting computer vision and deep learning-based geospatial solutions at scale. It is the first end-to-end geospatial machine learning platform, simplifying access to insights locked behind petabytes of imagery. Backed by a rigorous research methodology, this geospatial framework empowers researchers with powerful abstractions for image sourcing, dataset development, model development, large scale training, and model deployment. In this paper we provide the GeoEngine architecture explaining our design rationale in detail. We provide several real-world use cases of image sourcing, dataset development, and model building that have helped different organisations build and deploy geospatial solutions.
****
Revisiting Document Image Dewarping by Grid Regularization
Xiangwei Jiang, Rujiao Long, Nan Xue, Zhibo Yang, Cong Yao, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4543-4552
This paper addresses the problem of document image dewarping, which aims at eliminating the geometric distortion in document images for document digitization. Instead of designing a better neural network to approximate the optical flow fields between the inputs and outputs, we pursue the best readability by taking the text lines and the document boundaries into account from a constrained optimization perspective. Specifically, our proposed method first learns the boundary points and the pixels in the text lines and then follows the most simple observation that the boundaries and text lines in both horizontal and vertical directions should be kept after dewarping to introduce a novel grid regularization scheme. To obtain the final forward mapping for dewarping, we solve an optimization problem with our proposed grid regularization. The experiments comprehensively demonstrate that our proposed approach outperforms the prior arts by large margins in terms of readability (with the metrics of Character Errors Rate and the Edit Distance) while maintaining the best image quality on the publicly-available DocUNet benchmark.

****
Semi-Supervised Few-Shot Learning via Multi-Factor Clustering
Jie Ling, Lei Liao, Meng Yang, Jia Shuai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14564-14573
The scarcity of labeled data and the problem of model overfitting have been the challenges in few-shot learning. Recently, semi-supervised few-shot learning has been developed to obtain pseudo-labels of unlabeled samples for expanding the support set. However, the relationship between unlabeled and labeled data is not well exploited in generating pseudo labels, the noise of which will directly harm the model learning. In this paper, we propose a Clustering-based semi-supervised Few-Shot Learning (cluster-FSL) method to solve the above problems in image classification. By using multi-factor collaborative representation, a novel Multi-Factor Clustering (MFC) is designed to fuse the information of few-shot data distribution, which can generate soft and hard pseudo-labels for unlabeled samples based on labeled data. And we exploit the pseudo labels of unlabeled samples by MFC to expand the support set for obtaining more distribution information. Furthermore, robust data augmentation is used for support set in fine-tuning phase to increase the diversity of labeled samples. We verified the validity of the cluster-FSL by comparing it with other few-shot learning methods on three popular benchmark datasets, miniImageNet, tieredImageNet, and CUB-200-2011. The ablation experiments further demonstrate that our MFC can effectively fuse distribution information of labeled samples and provide high-quality pseudo-labels.
****
CMT-DeepLab: Clustering Mask Transformers for Panoptic Segmentation
Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2560-2570
We propose Clustering Mask Transformer (CMT-DeepLab), a transformer-based framework for panoptic segmentation designed around clustering. It rethinks the existing transformer architectures used in segmentation and detection; CMT-DeepLab considers the object queries as cluster centers, which fill the role of grouping the pixels when applied to segmentation. The clustering is computed with an alternating procedure, by first assigning pixels to the clusters by their feature affinity, and then updating the cluster centers and pixel features. Together, these operations comprise the Clustering Mask Transformer (CMT) layer, which produces cross-attention that is denser and more consistent with the final segmentation task. CMT-DeepLab improves the performance over prior art significantly by 4.4% PQ, achieving a new state-of-the-art of 55.7% PQ on the COCO test-dev set.
****
Weakly-Supervised Generation and Grounding of Visual Descriptions With Conditional Generative Models
Effrosyni Mavroudi, René Vidal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15544-15554
Given weak supervision from image- or video-caption pairs, we address the problem of grounding (localizing) each object word of a ground-truth or generated sentence describing a visual input. Recent weakly-supervised approaches leverage region proposals and ground words based on the region attention coefficients of captioning models. To predict each next word in the sentence they attend over regions using a summary of the previous words as a query, and then ground the word by selecting the most attended regions. However, this leads to sub-optimal grounding, since attention coefficients are computed without taking into account the word that needs to be localized. To address this shortcoming, we propose a novel Grounded Visual Description Conditional Variational Autoencoder (GVD-CVAE) and leverage its latent variables for grounding. In particular, we introduce a discrete random variable that models each word-to-region alignment, and learn its approximate posterior distribution given the full sentence. Experiments on challenging image and video datasets (Flickr30k Entities, YouCook2, ActivityNet Entities) validate the effectiveness of our conditional generative model, showing that it can substantially outperform soft-attention-based baselines in grounding.
****

Novel Class Discovery in Semantic Segmentation
Yuyang Zhao, Zhun Zhong, Nicu Sebe, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4340-4349
We introduce a new setting of Novel Class Discovery in Semantic Segmentation (NCDSS), which aims at segmenting unlabeled images containing new classes given prior knowledge from a labeled set of disjoint classes. In contrast to existing approaches that look at novel class discovery in image classification, we focus on the more challenging semantic segmentation. In NCDSS, we need to distinguish the objects and background, and to handle the existence of multiple classes within an image, which increases the difficulty in using the unlabeled data. To tackle this new setting, we leverage the labeled base data and a saliency model to coarsely cluster novel classes for model training in our basic framework. Additionally, we propose the Entropy-based Uncertainty Modeling and Self-training (EUMS) framework to overcome noisy pseudo-labels, further improving the model performance on the novel classes. Our EUMS utilizes an entropy ranking technique and a dynamic reassignment to distill clean labels, thereby making full use of the noisy data via self-supervised learning. We build the NCDSS benchmark on the PASCAL-5^i dataset and COCO-20^i dataset. Extensive experiments demonstrate the feasibility of the basic framework (achieving an average mIoU of 49.81% on PASCAL-5^i) and the effectiveness of EUMS framework (outperforming the basic framework by 9.28% mIoU on PASCAL-5^i).
****
ARCS: Accurate Rotation and Correspondence Search
Liangzu Peng, Manolis C. Tsakiris, René Vidal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11153-11163
This paper is about the old Wahba problem in its more general form, which we call "simultaneous rotation and correspondence search". In this generalization we need to find a rotation that best aligns two partially overlapping 3D point sets, of sizes m and n respectively with m\geq n. We first propose a solver, \texttt ARCS , that i) assumes noiseless point sets in general position, ii) requires only 2 inliers, iii) uses O(m\log m) time and O(m) space, and iv) can successfully solve the problem even with, e.g., m,n~ 10^6 in about 0.1 seconds. We next robustify \texttt ARCS to noise, for which we approximately solve consensus maximization problems using ideas from robust subspace learning and interval stabbing. Thirdly, we refine the approximately found consensus set by a Riemannian subgradient descent approach over the space of unit quaternions, which we show converges globally to an \varepsilon-stationary point in O(\varepsilon^ -4) iterations, or locally to the ground-truth at a linear rate in the absence of noise. We combine these algorithms into \texttt ARCS+ , to simultaneously search for rotations and correspondences. Experiments show that \texttt ARCS+ achieves state-of-the-art performance on large-scale datasets with more than 10^6 points with a 10^4 time-speedup over alternative methods. https://github.com/liangzu/ARCS
****
Learning To Anticipate Future With Dynamic Context Removal
Xinyu Xu, Yong-Lu Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12734-12744
Anticipating future events is an essential feature for intelligent systems and embodied AI. However, compared to the traditional recognition task, the uncertainty of future and reasoning ability requirement make the anticipation task very challenging and far beyond solved. In this filed, previous methods usually care more about the model architecture design or but few attention has been put on how to train an anticipation model with a proper learning policy. To this end, in this work, we propose a novel training scheme called Dynamic Context Removal (DCR), which dynamically schedule the visibility of observed future in the learning procedure. It follows the human-like curriculum learning process, i.e., gradually removing the event context to increase the anticipation difficulty till satisfying the final anticipation target. Our learning scheme is plug-and-play and easy to integrate any reasoning model including transformer and LSTM, with advantages in both effectiveness and efficiency. In extensive experiments, the proposed method achieves state-of-the-art on four widely-used benchmarks. Our code and mo

dels are publicly released at https://github.com/AllenXuuu/DCR.
****
GCFSR: A Generative and Controllable Face Super Resolution Method Without Facial and GAN Priors
Jingwen He, Wu Shi, Kai Chen, Lean Fu, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1889-1898
Face image super resolution (face hallucination) usually relies on facial priors to restore realistic details and preserve identity information. Recent advances can achieve impressive results with the help of GAN prior. They either design complicated modules to modify the fixed GAN prior or adopt complex training strategies to finetune the generator. In this work, we propose a generative and controllable face SR framework, called GCFSR, which can reconstruct images with faithful identity information without any additional priors. Generally, GCFSR has an encoder-generator architecture. Two modules called style modulation and feature modulation are designed for the multi-factor SR task. The style modulation aims to generate realistic face details and the feature modulation dynamically fuses the multi-level encoded features and the generated ones conditioned on the upscaling factor. The simple and elegant architecture can be trained from scratch in an end-to-end manner. For small upscaling factors (\leq8), GCFSR can produce surprisingly good results with only adversarial loss. After adding L1 and perceptual losses, GCFSR can outperform state-of-the-art methods for large upscaling factors (16, 32, 64). During the test phase, we can modulate the generative strength via feature modulation by changing the conditional upscaling factor continuously to achieve various generative effects. Code is available at https: //github.com/hejingwenhejingwen/GCFSR.
****
Perception Prioritized Training of Diffusion Models
Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11472-11481
Diffusion models learn to restore noisy data, which is corrupted with different levels of noise, by optimizing the weighted sum of the corresponding loss terms, i.e., denoising score matching loss. In this paper, we show that restoring data corrupted with certain noise levels offers a proper pretext task for the model to learn rich visual concepts. We propose to prioritize such noise levels over other levels during training, by redesigning the weighting scheme of the objective function. We show that our simple redesign of the weighting scheme significantly improves the performance of diffusion models regardless of the datasets, architectures, and sampling strategies.
****
Using 3D Topological Connectivity for Ghost Particle Reduction in Flow Reconstruction
Christina Tsalicoglou, Thomas Rösgen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1839-1847
Volumetric flow velocimetry for experimental fluid dynamics relies primarily on the 3D reconstruction of point objects, which are the detected positions of tracer particles identified in images obtained by a multi-camera setup. By assuming that the particles accurately follow the observed flow, their displacement over a known time interval is a measure of the local flow velocity. The number of particles imaged in a 1 Megapixel image is typically in the order of 1e3-1e4, resulting in a large number of consistent but incorrect reconstructions (no real particle in 3D), that must be eliminated through tracking or intensity constraints. In an alternative method, 3D Particle Streak Velocimetry (3D-PSV), the exposure time is increased, and the particles' pathlines are imaged as "streaks". We treat these streaks (a) as connected endpoints and (b) as conic section segments and develop a theoretical model that describes the mechanisms of 3D ambiguity generation and shows that streaks can drastically reduce reconstruction ambiguities. Moreover, we propose a method for simultaneously estimating these short, low-curvature conic section segments and their 3D position from multiple camera views. Our results validate the theory, and the streak and conic section reconstruction

method produces far fewer ambiguities than simple particle reconstruction, outperforming current state-of-the-art particle tracking software on the evaluated cases.
****

On the Integration of Self-Attention and Convolution
Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 815-825
Convolution and self-attention are two powerful techniques for representation learning, and they are usually considered as two peer approaches that are distinct from each other. In this paper, we show that there exists a strong underlying relation between them, in the sense that the bulk of computations of these two paradigms are in fact done with the same operation. Specifically, we first show that a traditional convolution with kernel size k x k can be decomposed into k^2 individual 1x1 convolutions, followed by shift and summation operations. Then, we interpret the projections of queries, keys, and values in selfattention module as multiple 1x1 convolutions, followed by the computation of attention weights and aggregation of the values. Therefore, the first stage of both two modules comprises the similar operation. More importantly, the first stage contributes a dominant computation complexity (square of the channel size) comparing to the second stage. This observation naturally leads to an elegant integration of these two seemingly distinct paradigms, i.e., a mixed model that enjoys the benefit of both self-Attention and Convolution (ACmix), while having minimum computational overhead compared to the pure convolution or selfattention counterpart. Extensive experiments show that our model achieves consistently improved results over competitive baselines on image recognition and downstream tasks. Code and pre-trained models will be released at https://github.com/LeapLabTHU/ACmix and https://gitee.com/mindspore/models.
****

Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction
Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, Guiqing Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6437-6446
This paper presents a high-quality human motion prediction method that accurately predicts future human poses given observed ones. Our method is based on the observation that a good initial guess of the future poses is very helpful in improving the forecasting accuracy. This motivates us to propose a novel two-stage prediction framework, including an init-prediction network that just computes the good guess and then a formal-prediction network that predicts the target future poses based on the guess. More importantly, we extend this idea further and design a multi-stage prediction framework where each stage predicts initial guess for the next stage, which brings more performance gain. To fulfill the prediction task at each stage, we propose a network comprising Spatial Dense Graph Convolutional Networks (S-DGCN) and Temporal Dense Graph Convolutional Networks (T-DGCN). Alternatively executing the two networks helps extract spatiotemporal features over the global receptive field of the whole pose sequence. All the above design choices cooperating together make our method outperform previous approaches by large margins: 6%-7% on Human3.6M, 5%-10% on CMU-MoCap, and 13%-16% on 3DPW.
****

CHEX: CHannel EXploration for CNN Model Compression
Zejiang Hou, Minghai Qin, Fei Sun, Xiaolong Ma, Kun Yuan, Yi Xu, Yen-Kuang Chen, Rong Jin, Yuan Xie, Sun-Yuan Kung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12287-12298
Channel pruning has been broadly recognized as an effective technique to reduce the computation and memory cost of deep convolutional neural networks. However, conventional pruning methods have limitations in that: they are restricted to pruning process only, and they require a fully pre-trained large model. Such limitations may lead to sub-optimal model quality as well as excessive memory and training cost. In this paper, we propose a novel Channel Exploration methodology, d

ubbed as CHEX, to rectify these problems. As opposed to pruning-only strategy, we propose to repeatedly prune and regrow the channels throughout the training process, which reduces the risk of pruning important channels prematurely. More exactly: From intra-layer's aspect, we tackle the channel pruning problem via a well-known column subset selection (CSS) formulation. From inter-layer's aspect, our regrowing stages open a path for dynamically re-allocating the number of channels across all the layers under a global channel sparsity constraint. In addition, all the exploration process is done in a single training from scratch without the need of a pre-trained large model. Experimental results demonstrate that CHEX can effectively reduce the FLOPs of diverse CNN architectures on a variety of computer vision tasks, including image classification, object detection, instance segmentation, and 3D vision. For example, our compressed ResNet-50 model on ImageNet dataset achieves 76% top-1 accuracy with only 25% FLOPs of the original ResNet-50 model, outperforming previous state-of-the-art channel pruning methods. The checkpoints and code are available at here.

****

M2I: From Factored Marginal Trajectory Prediction to Interactive Prediction

Qiao Sun, Xin Huang, Junru Gu, Brian C. Williams, Hang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6543-6552

Predicting future motions of road participants is an important task for driving autonomously in urban scenes. Existing models excel at predicting marginal trajectories for single agents, yet it remains an open question to jointly predict scene compliant trajectories over multiple agents. The challenge is due to exponentially increasing prediction space as a function of the number of agents. In this work, we exploit the underlying relations between interacting agents and decouple the joint prediction problem into marginal prediction problems. Our proposed approach M2I first classifies interacting agents as pairs of influencers and reactors, and then leverages a marginal prediction model and a conditional prediction model to predict trajectories for the influencers and reactors, respectively. The predictions from interacting agents are combined and selected according to their joint likelihoods. Experiments show that our simple but effective approach achieves state-of-the-art performance on the Waymo Open Motion Dataset interactive prediction benchmark.

****

Domain Adaptation on Point Clouds via Geometry-Aware Implicits

Yuefan Shen, Yanchao Yang, Mi Yan, He Wang, Youyi Zheng, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7223-7232

As a popular geometric representation, point clouds have attracted much attention in 3D vision, leading to many applications in autonomous driving and robotics. One important yet unsolved issue for learning on point cloud is that point clouds of the same object can have significant geometric variations if generated using different procedures or captured using different sensors. These inconsistencies induce domain gaps such that neural networks trained on one domain may fail to generalize on others. A typical technique to reduce the domain gap is to perform adversarial training so that point clouds in the feature space can align. However, adversarial training is easy to fall into degenerated local minima, resulting in negative adaptation gains. Here we propose a simple yet effective method for unsupervised domain adaptation on point clouds by employing a self-supervised task of learning geometry-aware implicits, which plays two critical roles in one shot. First, the geometric information in the point clouds is preserved through the implicit representations for downstream tasks. More importantly, the domain-specific variations can be effectively learned away in the implicit space. We also propose an adaptive strategy to compute unsigned distance fields for arbitrary point clouds due to the lack of shape models in practice. When combined with a task loss, the proposed outperforms state-of-the-art unsupervised domain adaptation methods that rely on adversarial domain alignment and more complicated self-supervised tasks. Our method is evaluated on both PointDA-10 and GraspNet datasets. Code and data are available at: https://github.com/Jhonve/ImplicitPCDA

****
Consistency Driven Sequential Transformers Attention Model for Partially Observable Scenes
Samrudhdhi B. Rangrej, Chetan L. Srinidhi, James J. Clark; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2518-2527
Most hard attention models initially observe a complete scene to locate and sense informative glimpses, and predict class-label of a scene based on glimpses. However, in many applications (e.g., aerial imaging), observing an entire scene is not always feasible due to the limited time and resources available for acquisition. In this paper, we develop a Sequential Transformers Attention Model (STAM) that only partially observes a complete image and predicts informative glimpse locations solely based on past glimpses. We design our agent using DeiT-distilled and train it with a one-step actor-critic algorithm. Furthermore, to improve classification performance, we introduce a novel training objective, which enforces consistency between the class distribution predicted by a teacher model from a complete image and the class distribution predicted by our agent using glimpses. When the agent senses only 4% of the total image area, the inclusion of the proposed consistency loss in our training objective yields 3% and 8% higher accuracy on ImageNet and fMoW datasets, respectively. Moreover, our agent outperforms previous state-of-the-art by observing nearly 27% and 42% fewer pixels in glimpses on ImageNet and fMoW.
****
GroupViT: Semantic Segmentation Emerges From Text Supervision
Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18134-18144
Grouping and recognition are important components of visual scene understanding, e.g., for object detection and semantic segmentation. With end-to-end deep learning systems, grouping of image regions usually happens implicitly via top-down supervision from pixel-level recognition labels. Instead, in this paper, we propose to bring back the grouping mechanism into deep networks, which allows semantic segments to emerge automatically with only text supervision. We propose a hierarchical Grouping Vision Transformer (GroupViT), which goes beyond the regular grid structure representation and learns to group image regions into progressively larger arbitrary-shaped segments. We train GroupViT jointly with a text encoder on a large-scale image-text dataset via contrastive losses. With only text supervision and without any pixel-level annotations, GroupViT learns to group together semantic regions and successfully transfers to the task of semantic segmentation in a zero-shot manner, i.e., without any further fine-tuning. It achieves a zero-shot accuracy of 52.3% mIoU on the PASCAL VOC 2012 and 22.4% mIoU on PASCAL Context datasets, and performs competitively to state-of-the-art transfer-learning methods requiring greater levels of supervision. We open-source our code at https://github.com/NVlabs/GroupViT
****
NeuralHOFusion: Neural Volumetric Rendering Under Human-Object Interactions
Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6155-6165
4D modeling of human-object interactions is critical for numerous applications. However, efficient volumetric capture and rendering of complex interaction scenarios, especially from sparse inputs, remain challenging. In this paper, we propose NeuralHOFusion, a neural approach for volumetric human-object capture and rendering using sparse consumer RGBD sensors. It marries traditional non-rigid fusion with recent neural implicit modeling and blending advances, where the captured humans and objects are layer-wise disentangled. For geometry modeling, we propose a neural implicit inference scheme with non-rigid key-volume fusion, as well as a template-aid robust object tracking pipeline. Our scheme enables detailed and complete geometry generation under complex interactions and occlusions. Moreover, we introduce a layer-wise human-object texture rendering scheme, which com

bines volumetric and image-based rendering in both spatial and temporal domains to obtain photo-realistic results. Extensive experiments demonstrate the effectiveness and efficiency of our approach in synthesizing photo-realistic free-view results under complex human-object interactions.
****

Generalizable Human Pose Triangulation
Kristijan Bartol, David Bojani■, Tomislav Petkovi■, Tomislav Pribani■; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11028-11037
We address the problem of generalizability for multi-view 3D human pose estimation. The standard approach is to first detect 2D keypoints in images and then apply triangulation from multiple views. Even though the existing methods achieve remarkably accurate 3D pose estimation on public benchmarks, most of them are limited to a single spatial camera arrangement and their number. Several methods address this limitation but demonstrate significantly degraded performance on novel views. We propose a stochastic framework for human pose triangulation and demonstrate a superior generalization across different camera arrangements on two public datasets. In addition, we apply the same approach to the fundamental matrix estimation problem, showing that the proposed method can successfully apply to other computer vision problems. The stochastic framework achieves more than 8.8% improvement on the 3D pose estimation task, compared to the state-of-the-art, and more than 30% improvement for fundamental matrix estimation, compared to a standard algorithm.
****

DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation
Gwanghyun Kim, Taesung Kwon, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2426-2435
Recently, GAN inversion methods combined with Contrastive Language-Image Pretraining (CLIP) enables zero-shot image manipulation guided by text prompts. However, their applications to diverse real images are still difficult due to the limited GAN inversion capability. Specifically, these approaches often have difficulties in reconstructing images with novel poses, views, and highly variable contents compared to the training data, altering object identity, or producing unwanted image artifacts. To mitigate these problems and enable faithful manipulation of real images, we propose a novel method, dubbed DiffusionCLIP, that performs text-driven image manipulation using diffusion models. Based on full inversion capability and high-quality image generation power of recent diffusion models, our method performs zero-shot image manipulation successfully even between unseen domains and takes another step towards general application by manipulating images from a widely varying ImageNet dataset. Furthermore, we propose a novel noise combination method that allows straightforward multi-attribute manipulation. Extensive experiments and human evaluation confirmed robust and superior manipulation performance of our methods compared to the existing baselines. Code is available at https://github.com/gwang-kim/DiffusionCLIP.git
****

Occlusion-Aware Cost Constructor for Light Field Depth Estimation
Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19809-19818
Matching cost construction is a key step in light field (LF) depth estimation, but was rarely studied in the deep learning era. Recent deep learning-based LF depth estimation methods construct matching cost by sequentially shifting each sub-aperture image (SAI) with a series of predefined offsets, which is complex and time-consuming. In this paper, we propose a simple and fast cost constructor to construct matching cost for LF depth estimation. Our cost constructor is composed by a series of convolutions with specifically designed dilation rates. By applying our cost constructor to SAI arrays, pixels under predefined disparities can be integrated and matching cost can be constructed without using any shifting operation. More importantly, the proposed cost constructor is occlusion-aware and can handle occlusions by dynamically modulating pixels from different views. Ba

sed on the proposed cost constructor, we develop a deep network for LF depth est imation. Our network ranks first on the commonly used 4D LF benchmark in terms o f the mean square error (MSE), and achieves a faster running time than other sta te-of-the-art methods.
****

SmartPortraits: Depth Powered Handheld Smartphone Dataset of Human Portraits for State Estimation, Reconstruction and Synthesis
Anastasiia Kornilova, Marsel Faizullin, Konstantin Pakulev, Andrey Sadkov, Denis Kukushkin, Azat Akhmetyanov, Timur Akhtyamov, Hekmat Taherinejad, Gonzalo Ferre r; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2022, pp. 21318-21329
We present a dataset of 1000 video sequences of human portraits recorded in real and uncontrolled conditions by using a handheld smartphone accompanied by an ex ternal high-quality depth camera. The collected dataset contains 200 people capt ured in different poses and locations and its main purpose is to bridge the gap between raw measurements obtained from a smartphone and downstream applications, such as state estimation, 3D reconstruction, view synthesis, etc. The sensors e mployed in data collection are the smartphone's camera and Inertial Measurement Unit (IMU), and an external Azure Kinect DK depth camera software synchronized w ith sub-millisecond precision to the smartphone system. During the recording, th e smartphone flash is used to provide a periodic secondary source of lightning. Accurate mask of the foremost person is provided as well as its impact on camera alignment accuracy. For evaluation purposes, we compare multiple state-of-the-a rt camera alignment methods by using a Motion Capture system. We provide a smart phone visual-inertial benchmark for portrait capturing, where we report results for multiple methods and motivate further use of the provided trajectories, avai lable in the dataset, in view synthesis and 3D reconstruction tasks.
****

BppAttack: Stealthy and Efficient Trojan Attacks Against Deep Neural Networks vi a Image Quantization and Contrastive Adversarial Learning
Zhenting Wang, Juan Zhai, Shiqing Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15074-15084
Deep neural networks are vulnerable to Trojan attacks. Existing attacks use visi ble patterns (e.g., a patch or image transformations) as triggers, which are vul nerable to human inspection. In this paper, we propose stealthy and efficient Tr ojan attacks, BppAttack. Based on existing biology literature on human visual sy stems, we propose to use image quantization and dithering as the Trojan trigger, making imperceptible changes. It is a stealthy and efficient attack without tra ining auxiliary models. Due to the small changes made to images, it is hard to i nject such triggers during training. To alleviate this problem, we propose a con trastive learning based approach that leverages adversarial attacks to generate negative sample pairs so that the learned trigger is precise and accurate. The p roposed method achieves high attack success rates on four benchmark datasets, in cluding MNIST, CIFAR-10, GTSRB, and CelebA. It also effectively bypasses existin g Trojan defenses and human inspection. Our code can be found in https://github. com/RU-System-Software-and-Security/BppAttack.
****

GlideNet: Global, Local and Intrinsic Based Dense Embedding NETwork for Multi-Ca tegory Attributes Prediction
Kareem Metwaly, Aerin Kim, Elliot Branson, Vishal Monga; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 483 5-4846
Attaching attributes (such as color, shape, state, action) to object categories is an important computer vision problem. Attribute prediction has seen exciting recent progress and is often formulated as a multi-label classification problem. Yet significant challenges remain in: 1) predicting a large number of attribute s over multiple object categories, 2) modeling category-dependence of attributes , 3) methodically capturing both global and local scene context, and 4) robustly predicting attributes of objects with low pixel-count. To address these issues, we propose a novel multi-category attribute prediction deep architecture named

GlideNet, which contains three distinct feature extractors. A global feature extractor recognizes what objects are present in a scene, whereas a local one focuses on the area surrounding the object of interest. Meanwhile, an intrinsic feature extractor uses an extension of standard convolution dubbed Informed Convolution to retrieve features of objects with low pixel-count utilizing its binary mask. GlideNet then uses gating mechanisms with binary masks and its self-learned category embedding to combine the dense embeddings. Collectively, the Global-Local-Intrinsic blocks comprehend the scene's global context while attending to the characteristics of the local object of interest. The architecture adapts the feature composition based on the category via category embedding. Finally, using the combined features, an interpreter predicts the attributes, and the length of the output is determined by the category, thereby removing unnecessary attributes. GlideNet can achieve compelling results on two recent and challenging datasets -- VAW and CAR -- for large-scale attribute prediction. For instance, it obtains more than 5% gain over state of the art in the mean recall (mR) metric. GlideNet's advantages are especially apparent when predicting attributes of objects with low pixel counts as well as attributes that demand global context understanding. Finally, we show that GlideNet excels in training starved real-world scenarios.

****

Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation

Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, Liqiang Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19427-19436

Scene Graph Generation, which generally follows a regular encoder-decoder pipeline, aims to first encode the visual contents within the given image and then parse them into a compact summary graph. Existing SGG approaches generally not only neglect the insufficient modality fusion between vision and language, but also fail to provide informative predicates due to the biased relationship predictions, leading SGG far from practical. Towards this end, in this paper, we first present a novel Stacked Hybrid-Attention network, which facilitates the intra-modal refinement as well as the inter-modal interaction, to serve as the encoder. We then devise an innovative Group Collaborative Learning strategy to optimize the decoder. Particularly, based upon the observation that the recognition capability of one classifier is limited towards an extremely unbalanced dataset, we first deploy a group of classifiers that are expert in distinguishing different subsets of classes, and then cooperatively optimize them from two aspects to promote the unbiased SGG. Experiments conducted on VG and GQA datasets demonstrate that, we not only establish a new state-of-the-art in the unbiased metric, but also nearly double the performance compared with two baselines. Our code is available at https://github.com/dongxingning/SHA-GCL-for-SGG.

****

Ensembling Off-the-Shelf Models for GAN Training

Nupur Kumari, Richard Zhang, Eli Shechtman, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10651-10662

The advent of large-scale training has produced a cornucopia of powerful visual recognition models. However, generative models, such as GANs, have traditionally been trained from scratch in an unsupervised manner. Can the collective "knowledge" from a large bank of pretrained vision models be leveraged to improve GAN training? If so, with so many models to choose from, which one(s) should be selected, and in what manner are they most effective? We find that pretrained computer vision models can significantly improve performance when used in an ensemble of discriminators. Notably, the particular subset of selected models greatly affects performance. We propose an effective selection mechanism, by probing the linear separability between real and fake samples in pretrained model embeddings, choosing the most accurate model, and progressively adding it to the discriminator ensemble. Interestingly, our method can improve GAN training in both limited data and large-scale settings. Given only 10k training samples, our FID on LSUN C

at matches the StyleGAN2 trained on 1.6M images. On the full dataset, our method improves FID by 1.5 to 2 times on cat, church, and horse categories of LSUN.
****

Towards Better Plasticity-Stability Trade-Off in Incremental Learning: A Simple Linear Connector

Guoliang Lin, Hanlu Chu, Hanjiang Lai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 89-98

Plasticity-stability dilemma is a main problem for incremental learning, where plasticity is referring to the ability to learn new knowledge, and stability retains the knowledge of previous tasks. Many methods tackle this problem by storing previous samples, while in some applications, training data from previous tasks cannot be legally stored. In this work, we propose to employ mode connectivity in loss landscapes to achieve better plasticity-stability trade-off without any previous samples. We give an analysis of why and how to connect two independently optimized optima of networks, null-space projection for previous tasks and simple SGD for the current task, can attain a meaningful balance between preserving already learned knowledge and granting sufficient flexibility for learning a new task. This analysis of mode connectivity also provides us a new perspective and technology to control the trade-off between plasticity and stability. We evaluate the proposed method on several benchmark datasets. The results indicate our simple method can achieve notable improvement, and perform well on both the past and current tasks. On 10-split-CIFAR-100 task, our method achieves 79.79% accuracy, which is 6.02% higher. Our method also achieves 6.33% higher accuracy on TinyImageNet. Code is available at https://github.com/lingl1024/Connector.
****

Topology-Preserving Shape Reconstruction and Registration via Neural Diffeomorphic Flow

Shanlin Sun, Kun Han, Deying Kong, Hao Tang, Xiangyi Yan, Xiaohui Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20845-20855

Deep Implicit Functions (DIFs) represent 3D geometry with continuous signed distance functions learned through deep neural nets. Recently DIFs-based methods have been proposed to handle shape reconstruction and dense point correspondences simultaneously, capturing semantic relationships across shapes of the same class by learning a DIFs-modeled shape template. These methods provide great flexibility and accuracy in reconstructing 3D shapes and inferring correspondences. However, the point correspondences built from these methods do not intrinsically preserve the topology of the shapes, unlike mesh-based template matching methods. This limits their applications on 3D geometries where underlying topological structures exist and matter, such as anatomical structures in medical images. In this paper, we propose a new model called Neural Diffeomorphic Flow (NDF) to learn deep implicit shape templates, representing shapes as conditional diffeomorphic deformations of templates, intrinsically preserving shape topologies. The diffeomorphic deformation is realized by an auto-decoder consisting of Neural Ordinary Differential Equation (NODE) blocks that progressively map shapes to implicit templates. We conduct extensive experiments on several medical image organ segmentation datasets to evaluate the effectiveness of NDF on reconstructing and aligning shapes. NDF achieves consistently state-of-the-art organ shape reconstruction and registration results in both accuracy and quality. The source code is publicly available at https://github.com/Siwensun/Neural_Diffeomorphic_Flow--NDF.
****

Segment and Complete: Defending Object Detectors Against Adversarial Patch Attacks With Robust Patch Detection

Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, Soheil Feizi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14973-14982

Object detection plays a key role in many security-critical systems. Adversarial patch attacks, which are easy to implement in the physical world, pose a serious threat to state-of-the-art object detectors. Developing reliable defenses for object detectors against patch attacks is critical but severely understudied. In

this paper, we propose Segment and Complete defense (SAC), a general framework for defending object detectors against patch attacks through detection and removal of adversarial patches. We first train a patch segmenter that outputs patch masks which provide pixel-level localization of adversarial patches. We then propose a self adversarial training algorithm to robustify the patch segmenter. In addition, we design a robust shape completion algorithm, which is guaranteed to remove the entire patch from the images if the outputs of the patch segmenter are within a certain Hamming distance of the ground-truth patch masks. Our experiments on COCO and xView datasets demonstrate that SAC achieves superior robustness even under strong adaptive attacks with no reduction in performance on clean images, and generalizes well to unseen patch shapes, attack budgets, and unseen attack methods. Furthermore, we present the APRICOT-Mask dataset, which augments the APRICOT dataset with pixel-level annotations of adversarial patches. We show SAC can significantly reduce the targeted attack success rate of physical patch attacks. Our code is available at https://github.com/joellliu/SegmentAndComplete.

****

Cross-Domain Few-Shot Learning With Task-Specific Adapters

Wei-Hong Li, Xialei Liu, Hakan Bilen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7161-7170

In this paper, we look at the problem of cross-domain few-shot classification that aims to learn a classifier from previously unseen classes and domains with few labeled samples. Recent approaches broadly solve this problem by parameterizing their few-shot classifiers with task-agnostic and task-specific weights where the former is typically learned on a large training set and the latter is dynamically predicted through an auxiliary network conditioned on a small support set. In this work, we focus on the estimation of the latter, and propose to learn task-specific weights from scratch directly on a small support set, in contrast to dynamically estimating them. In particular, through systematic analysis, we show that task-specific weights through parametric adapters in matrix form with residual connections to multiple intermediate layers of a backbone network significantly improves the performance of the state-of-the-art models in the Meta-Dataset benchmark with minor additional cost.

****

MAXIM: Multi-Axis MLP for Image Processing

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, Yinxiao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5769-5780

Recent progress on Transformers and multi-layer perceptron (MLP) models provide new network architectural designs for computer vision tasks. Although these models proved to be effective in many vision tasks such as image recognition, there remain challenges in adapting them for low-level vision. The inflexibility to support high-resolution images and limitations of local attention are perhaps the main bottlenecks. In this work, we present a multi-axis MLP based architecture called MAXIM, that can serve as an efficient and flexible general-purpose vision backbone for image processing tasks. MAXIM uses a UNet-shaped hierarchical structure and supports long-range interactions enabled by spatially-gated MLPs. Specifically, MAXIM contains two MLP-based building blocks: a multi-axis gated MLP that allows for efficient and scalable spatial mixing of local and global visual cues, and a cross-gating block, an alternative to cross-attention, which accounts for cross-feature conditioning. Both these modules are exclusively based on MLPs, but also benefit from being both global and 'fully-convolutional', two properties that are desirable for image processing. Our extensive experimental results show that the proposed MAXIM model achieves state-of-the-art performance on more than ten benchmarks across a range of image processing tasks, including denoising, deblurring, deraining, dehazing, and enhancement while requiring fewer or comparable numbers of parameters and FLOPs than competitive models. The source code and trained models will be available at https://github.com/google-research/maxim.

****

Learning Part Segmentation Through Unsupervised Domain Adaptation From Synthetic Vehicles
Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, Alan Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19140-19151
Part segmentations provide a rich and detailed part-level description of objects. However, their annotation requires an enormous amount of work, which makes it difficult to apply standard deep learning methods. In this paper, we propose the idea of learning part segmentation through unsupervised domain adaptation (UDA) from synthetic data. We first introduce UDA-Part, a comprehensive part segmentation dataset for vehicles that can serve as an adequate benchmark for UDA (https://qliu24.github.io/udapart/). In UDA-Part, we label parts on 3D CAD models which enables us to generate a large set of annotated synthetic images. We also annotate parts on a number of real images to provide a real test set. Secondly, to advance the adaptation of part models trained from the synthetic data to the real images, we introduce a new UDA algorithm that leverages the object's spatial structure to guide the adaptation process. Our experimental results on two real test datasets confirm the superiority of our approach over existing works, and demonstrate the promise of learning part segmentation for general objects from synthetic data. We believe our dataset provides a rich testbed to study UDA for part segmentation and will help to significantly push forward research in this area.
****
Delving Into the Estimation Shift of Batch Normalization in a Network
Lei Huang, Yi Zhou, Tian Wang, Jie Luo, Xianglong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 763-772
Batch normalization (BN) is a milestone technique in deep learning. It normalizes the activation using mini-batch statistics during training but the estimated population statistics during inference. This paper focuses on investigating the estimation of population statistics. We define the estimation shift magnitude of BN to quantitatively measure the difference between its estimated population statistics and expected ones. Our primary observation is that the estimation shift can be accumulated due to the stack of BN in a network, which has detriment effects for the test performance. We further find a batch-free normalization (BFN) can block such an accumulation of estimation shift. These observations motivate our design of XBNBlock that replace one BN with BFN in the bottleneck block of residual-style networks. Experiments on the ImageNet and COCO benchmarks show that XBNBlock consistently improves the performance of different architectures, including ResNet and ResNeXt, by a significant margin and seems to be more robust to distribution shift.
****
Towards Better Understanding Attribution Methods
Sukrut Rao, Moritz Böhle, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10223-10232
Deep neural networks are very successful on many vision tasks, but hard to interpret due to their black box nature. To overcome this, various post-hoc attribution methods have been proposed to identify image regions most influential to the models' decisions. Evaluating such methods is challenging since no ground truth attributions exist. We thus propose three novel evaluation schemes to more reliably measure the faithfulness of those methods, to make comparisons between them more fair, and to make visual inspection more systematic. To address faithfulness, we propose a novel evaluation setting (DiFull) in which we carefully control which parts of the input can influence the output in order to distinguish possible from impossible attributions. To address fairness, we note that different methods are applied at different layers, which skews any comparison, and so evaluate all methods on the same layers (ML-Att) and discuss how this impacts their performance on quantitative metrics. For more systematic visualizations, we propose a scheme (AggAtt) to qualitatively evaluate the methods on complete datasets. We use these evaluation schemes to study strengths and shortcomings of some widel

y used attribution methods. Finally, we propose a post-processing smoothing step that significantly improves the performance of some attribution methods, and discuss its applicability.

****

Learning Object Context for Novel-View Scene Layout Generation
Xiaotian Qiao, Gerhard P. Hancke, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16990-16999

Novel-view prediction of a scene has many applications. Existing works mainly focus on generating novel-view images via pixel-wise prediction in the image space, often resulting in severe ghosting and blurry artifacts. In this paper, we make the first attempt to explore novel-view prediction in the layout space, and introduce the new problem of novel-view scene layout generation. Given a single scene layout and the camera transformation as inputs, our goal is to generate a plausible scene layout for a specified viewpoint. Such a problem is challenging as it involves accurate understanding of the 3D geometry and semantics of the scene from as little as a single 2D scene layout. To tackle this challenging problem, we propose a deep model to capture contextualized object representation by explicitly modeling the object context transformation in the scene. The contextualized object representation is essential in generating geometrically and semantically consistent scene layouts of different views. Experiments show that our model outperforms several strong baselines on many indoor and outdoor scenes, both qualitatively and quantitatively. We also show that our model enables a wide range of applications, including novel-view image synthesis, novel-view image editing, and amodal object estimation.

****

PSTR: End-to-End One-Step Person Search With Transformers
Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9458-9467

We propose a novel one-step transformer-based person search framework, PSTR, that jointly performs person detection and re-identification (re-id) in a single architecture. PSTR comprises a person search-specialized (PSS) module that contains a detection encoder-decoder for person detection along with a discriminative re-id decoder for person re-id. The discriminative re-id decoder utilizes a multi-level supervision scheme with a shared decoder for discriminative re-id feature learning and also comprises a part attention block to encode relationship between different parts of a person. We further introduce a simple multi-scale scheme to support re-id across person instances at different scales. PSTR jointly achieves the diverse objectives of object-level recognition (detection) and instance-level matching (re-id). To the best of our knowledge, we are the first to propose an end-to-end one-step transformer-based person search framework. Experiments are performed on two popular benchmarks: CUHK-SYSU and PRW. Our extensive ablations reveal the merits of the proposed contributions. Further, the proposed PSTR sets a new state-of-the-art on both benchmarks. On the challenging PRW benchmark, PSTR achieves a mean average precision (mAP) score of 56.5%. The source code is available at https://github.com/JialeCao001/PSTR.

****

Neural Fields As Learnable Kernels for 3D Reconstruction
Francis Williams, Zan Gojcic, Sameh Khamis, Denis Zorin, Joan Bruna, Sanja Fidler, Or Litany; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18500-18510

We present Neural Kernel Fields: a novel method for reconstructing implicit 3D shapes based on a learned kernel ridge regression. Our technique achieves state-of-the-art results when reconstructing 3D objects and large scenes from sparse oriented points, and can reconstruct shape categories outside the training set with almost no drop in accuracy. The core insight of our approach is that kernel methods are extremely effective for reconstructing shapes when the chosen kernel has an appropriate inductive bias. We thus factor the problem of shape reconstruction into two parts: (1) a backbone neural network which learns kernel parameter

s from data, and (2) a kernel ridge regression that fits the input points on-the-fly by solving a simple positive definite linear system using the learned kernel. As a result of this factorization, our reconstruction gains the benefits of data-driven methods under sparse point density while maintaining interpolatory behavior, which converges to the ground truth shape as input sampling density increases. Our experiments demonstrate a strong generalization capability to objects outside the train-set category and scanned scenes.

****

A Deeper Dive Into What Deep Spatiotemporal Networks Encode: Quantifying Static vs. Dynamic Information

Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil D. B. Bruce, Richard P. Wildes, Konstantinos G. Derpanis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13999-14009

Deep spatiotemporal models are used in a variety of computer vision tasks, such as action recognition and video object segmentation. Currently, there is a limited understanding of what information is captured by these models in their intermediate representations. For example, while it has been observed that action recognition algorithms are heavily influenced by visual appearance in single static frames, there is no quantitative methodology for evaluating such static bias in the latent representation compared to bias toward dynamic information (e.g., motion). We tackle this challenge by proposing a novel approach for quantifying the static and dynamic biases of any spatiotemporal model. To show the efficacy of our approach, we analyse two widely studied tasks, action recognition and video object segmentation. Our key findings are threefold: (i) Most examined spatiotemporal models are biased toward static information; although, certain two-stream architectures with cross-connections show a better balance between the static and dynamic information captured. (ii) Some datasets that are commonly assumed to be biased toward dynamics are actually biased toward static information. (iii) Individual units (channels) in an architecture can be biased toward static, dynamic or a combination of the two.

****

Detector-Free Weakly Supervised Group Activity Recognition

Dongkeun Kim, Jinsung Lee, Minsu Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20083-20093

Group activity recognition is the task of understanding the activity conducted by a group of people as a whole in a multi-person video. Existing models for this task are often impractical in that they demand ground-truth bounding box labels of actors even in testing or rely on off-the-shelf object detectors. Motivated by this, we propose a novel model for group activity recognition that depends neither on bounding box labels nor on object detector. Our model based on Transformer localizes and encodes partial contexts of a group activity by leveraging the attention mechanism, and represents a video clip as a set of partial context embeddings. The embedding vectors are then aggregated to form a single group representation that reflects the entire context of an activity while capturing temporal evolution of each partial context. Our method achieves outstanding performance on two benchmarks, Volleyball and NBA datasets, surpassing not only the state of the art trained with the same level of supervision, but also some of existing models relying on stronger supervision.

****

NFormer: Robust Person Re-Identification With Neighbor Transformer

Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, Efstratios Gavves; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7297-7307

Person re-identification aims to retrieve persons in highly varying settings across different cameras and scenarios, in which robust and discriminative representation learning is crucial. Most research considers learning representations from single images, ignoring any potential interactions between them. However, due to the high intra-identity variations, ignoring such interactions typically leads to outlier features. To tackle this issue, we propose a Neighbor Transformer Network, or NFormer, which explicitly models interactions across all input images

, thus suppressing outlier features and leading to more robust representations o
verall. As modelling interactions between enormous amount of images is a massive
 task with lots of dis- tractors, NFormer introduces two novel modules, the Land
mark Agent Attention, and the Reciprocal Neighbor Softmax. Specifically, the Lan
dmark Agent Attention efficiently models the relation map between images by a lo
w-rank factorization with a few landmarks in feature space. Moreover, the Recipr
ocal Neighbor Softmax achieves sparse attention to relevant -rather than all- ne
ighbors only, which alleviates interference of irrelevant representations and fu
rther relieves the computational burden. In experiments on four large-scale data
sets, NFormer achieves a new state-of-the-art. The code is released at https://g
ithub.com/haochenheheda/NFormer.
****

Joint Forecasting of Panoptic Segmentations With Difference Attention
Colin Graber, Cyril Jazra, Wenjie Luo, Liangyan Gui, Alexander G. Schwing; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2022, pp. 2627-2636
Forecasting of a representation is important for safe and effective autonomy. Fo
r this, panoptic segmentations have been studied as a compelling representation
 in recent work. However, recent state-of-the-art on panoptic segmentation foreca
sting suffers from two issues: first, individual object instances are treated in
dependently of each other; second, individual object instance forecasts are merg
ed in a heuristic manner. To address both issues, we study a new panoptic segmen
tation forecasting model that jointly forecasts all object instances in a scene
 using a transformer model based on 'difference attention.' It further refines th
e predictions by taking depth estimates into account. We evaluate the proposed m
odel on the Cityscapes and AIODrive datasets. We find difference attention to be
 particularly suitable for forecasting because the difference of quantities like
 locations enables a model to explicitly reason about velocities and acceleratio
n. Because of this, we attain state-of-the-art on panoptic segmentation forecast
ing metrics.
****

HairCLIP: Design Your Hair by Text and Reference Image
Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming
Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2022, pp. 18072-18081
Hair editing is an interesting and challenging problem in computer vision and gr
aphics. Many existing methods require well-drawn sketches or masks as conditiona
l inputs for editing, however these interactions are neither straightforward nor
 efficient. In order to free users from the tedious interaction process, this pa
per proposes a new hair editing interaction mode, which enables manipulating hai
r attributes individually or jointly based on the texts or reference images prov
ided by users. For this purpose, we encode the image and text conditions in a sh
ared embedding space and propose a unified hair editing framework by leveraging
 the powerful image text representation capability of the Contrastive Language-Im
age Pre-Training (CLIP) model. With the carefully designed network structures an
d loss functions, our framework can perform high-quality hair editing in a disen
tangled manner. Extensive experiments demonstrate the superiority of our approac
h in terms of manipulation accuracy, visual realism of editing results, and irre
levant attribute preservation.
****

Imposing Consistency for Optical Flow Estimation
Jisoo Jeong, Jamie Menjay Lin, Fatih Porikli, Nojun Kwak; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 31
81-3191
Imposing consistency through proxy tasks has been shown to enhance data-driven l
earning and enable self-supervision in various tasks. This paper introduces nove
l and effective consistency strategies for optical flow estimation, a problem wh
ere labels from real-world data are very challenging to derive. More specificall
y, we propose occlusion consistency and zero forcing in the forms of self-superv
ised learning and transformation consistency in the form of semi-supervised lear

ning. We apply these consistency techniques in a way that the network model learns to describe pixel-level motions better while requiring no additional annotations. We demonstrate that our consistency strategies applied to a strong baseline network model using the original datasets and labels provide further improvements, attaining the state-of-the-art results on the KITTI-2015 scene flow benchmark in the non-stereo category. Our method achieves the best foreground accuracy (4.33% in Fl-all) over both the stereo and non-stereo categories, even though using only monocular image inputs.
****

Style Transformer for Image Inversion and Editing
Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, Qingli Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11337-11346
Existing GAN inversion methods fail to provide codes for reliable reconstruction and flexible editing simultaneously. This paper presents a transformer-based image inversion and editing model for pretrained StyleGAN which is not only with less distortions, but also of high quality and flexibility for editing. The proposed model employs a CNN encoder to provide multi-scale image features as keys and values. Meanwhile it regards the style code to be determined for different layers of the generator as queries. It first initializes query tokens as learnable parameters and maps them into W+ space. Then the multi-stage alternate self- and cross-attention are utilized, updating queries with the purpose of inverting the input by the generator. Moreover, based on the inverted code, we investigate the reference- and label-based attribute editing through a pretrained latent classifier, and achieve flexible image-to-image translation with high quality results. Extensive experiments are carried out, showing better performances on both inversion and editing tasks within StyleGAN.
****

OakInk: A Large-Scale Knowledge Repository for Understanding Hand-Object Interaction
Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20953-20962
Learning how humans manipulate objects requires machines to acquire knowledge from two perspectives: one for understanding object affordances and the other for learning human's interactions based on the affordances. Even though these two knowledge bases are crucial, we find that current databases lack a comprehensive awareness of them. In this work, we propose a multi-modal and rich-annotated knowledge repository, OakInk, for visual and cognitive understanding of hand-object interactions. We start to collect 1,800 common household objects and annotate their affordances to construct the first knowledge base: Oak. Given the affordance, we record rich human interactions with 100 selected objects in Oak. Finally, we transfer the interactions on the 100 recorded objects to their virtual counterparts through a novel method: Tink. The recorded and transferred hand-object interactions constitute the second knowledge base: Ink. As a result, OakInk contains 50,000 distinct affordance-aware and intent-oriented hand-object interactions. We benchmark OakInk on pose estimation and grasp generation tasks. Moreover, we propose two practical applications of OakInk: intent-based interaction generation and handover generation. Our dataset and source code are publicly available at www.oakink.net.
****

Pyramid Adversarial Training Improves ViT Performance
Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, Deqing Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13419-13429
Aggressive data augmentation is a key component of the strong generalization capabilities of Vision Transformer (ViT). One such data augmentation technique is adversarial training (AT); however, many prior works have shown that this often results in poor clean accuracy. In this work, we present pyramid adversarial training (PyramidAT), a simple and effective technique to improve ViT's overall perf

ormance. We pair it with a "matched" Dropout and stochastic depth regularization , which adopts the same Dropout and stochastic depth configuration for the clean and adversarial samples. Similar to the improvements on CNNs by AdvProp (not dir ectly applicable to ViT), our pyramid adversarial training breaks the trade-off between in-distribution accuracy and out-of-distribution robustness for ViT and related architectures. It leads to 1.82% absolute improvement on ImageNet clean accuracy for the ViT-B model when trained only on ImageNet-1K data, while simult aneously boosting performance on 7 ImageNet robustness metrics, by absolute numb ers ranging from 1.76% to 15.68%. We set a new state-of-the-art for ImageNet-C (41.42 mCE), ImageNet-R (53.92%), and ImageNet-Sketch (41.04%) without extra data , using only the ViT-B/16 backbone and our pyramid adversarial training. Our cod e is publicly available at pyramidat.github.io.

****

Bridging Global Context Interactions for High-Fidelity Image Completion

Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, Dinh Phung; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11512 -11522

Bridging global context interactions correctly is important for high-fidelity im age completion with large masks. Previous methods attempting this via deep or la rge receptive field (RF) convolutions cannot escape from the dominance of nearby interactions, which may be inferior. In this paper, we propose to treat image c ompletion as a directionless sequence-to-sequence prediction task, and deploy a transformer to directly capture long-range dependence. Crucially, we employ a re strictive CNN with small and non-overlapping RF for weighted token representatio n, which allows the transformer to explicitly model the long-range visible conte xt relations with equal importance in all layers, without implicitly confounding neighboring tokens when larger RFs are used. To improve appearance consistency between visible and generated regions, a novel attention-aware layer (AAL) is in troduced to better exploit distantly related high-frequency features. Overall, e xtensive experiments demonstrate superior performance compared to state-of-the-a rt methods on several datasets. Code is available at https://github.com/lyndonzh eng/TFill.

****

SwinBERT: End-to-End Transformers With Sparse Attention for Video Captioning

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, Lijuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17949-17958

The canonical approach to video captioning dictates a caption generation model t o learn from offline-extracted dense video features. These feature extractors us ually operate on video frames sampled at a fixed frame rate and are often traine d on image/video understanding tasks, without adaption to video captioning data. In this work, we present SwinBERT, an end-to-end transformer-based model for vi deo captioning, which takes video frame patches directly as inputs, and outputs a natural language description. Instead of leveraging multiple 2D/3D feature ext ractors, our method adopts a video transformer to encode spatial-temporal repres entations that can adapt to variable lengths of video input without dedicated de sign for different frame rates. Based on this model architecture, we show that v ideo captioning can benefit significantly from more densely sampled video frames as opposed to previous successes with sparsely sampled video frames for video-a nd-language understanding tasks (e.g., video question answering). Moreover, to a void the inherent redundancy in consecutive video frames, we propose adaptively learning a sparse attention mask and optimizing it for task-specific performance improvement through better long-range video sequence modeling. Through extensiv e experiments on 5 video captioning datasets, we show that SwinBERT achieves acr oss-the-board performance improvements over previous methods, often by a large m argin. The learned sparse attention masks in addition push the limit to new stat e of the arts, and can be transferred between different video lengths and betwee n different datasets. Code is available at https://github.com/microsoft/SwinBERT

****

Maximum Spatial Perturbation Consistency for Unpaired Image-to-Image Translation

Yanwu Xu, Shaoan Xie, Wenhao Wu, Kun Zhang, Mingming Gong, Kayhan Batmanghelich; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18311-18320

Unpaired image-to-image translation (I2I) is an ill-posed problem, as an infinite number of translation functions can map the source domain distribution to the target distribution. Therefore, much effort has been put into designing suitable constraints, e.g., cycle consistency (CycleGAN), geometry consistency (GCGAN), and contrastive learning-based constraints (CUTGAN), that help better pose the problem. However, these well-known constraints have limitations: (1) they are either too restrictive or too weak for specific I2I tasks; (2) these methods result in content distortion when there is a significant spatial variation between the source and target domains. This paper proposes a universal regularization technique called maximum spatial perturbation consistency (MSPC), which enforces a spatial perturbation function (T) and the translation operator (G) to be commutative (i.e., T \circ G = G \circ T). In addition, we introduce two adversarial training components for learning the spatial perturbation function. The first one lets T compete with G to achieve maximum perturbation. The second one lets G and T compete with discriminators to align the spatial variations caused by the change of object size, object distortion, background interruptions, etc. Our method outperforms the state-of-the-art methods on most I2I benchmarks. We also introduce a new benchmark, namely the front face to profile face dataset, to emphasize the underlying challenges of I2I for real-world applications. We finally perform ablation experiments to study the sensitivity of our method to the severity of spatial perturbation and its effectiveness for distribution alignment.

****

Unseen Classes at a Later Time? No Problem

Hari Chandana Kuchibhotla, Sumitra S Malagi, Shivam Chandhok, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9245-9254

Recent progress towards learning from limited supervision has encouraged efforts towards designing models that can recognize novel classes at test time (generalized zero-shot learning or GZSL). GZSL approaches assume knowledge of all classes, with or without labeled data, before-hand. However, practical scenarios demand models that are adaptable and can handle dynamic addition of new seen and unseen classes on the fly (i.e continual generalized zero-shot learning or CGZSL). One solution is to sequentially retrain and reuse conventional GZSL methods, however, such an approach suffers from catastrophic forgetting leading to suboptimal generalization performance. A few recent efforts towards tackling CGZSL have been limited by difference in settings, practicality, data splits and protocols followed - inhibiting fair comparison and a clear direction forward. Motivated from these observations, in this work, we firstly consolidate the different CGZSL setting variants and propose a new Online CGZSL setting which is more practical and flexible. Secondly, we introduce a unified feature-generative framework for CGZSL that leverages bi-directional incremental alignment to dynamically adapt to addition of new classes with or without labeled data that arrive over time in any of these CGZSL settings. Our comprehensive experiments and analysis on five benchmark datasets and comparison with baselines show that our approach consistently outperforms existing methods, especially on the more practical Online setting.

****

InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering

Mijeong Kim, Seonguk Seo, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12912-12921

We present an information-theoretic regularization technique for few-shot novel view synthesis based on neural implicit representation. The proposed approach minimizes potential reconstruction inconsistency that happens due to insufficient viewpoints by imposing the entropy constraint of the density in each ray. In addition, to alleviate the potential degenerate issue when all training images are acquired from almost redundant viewpoints, we further incorporate the spatially smoothness constraint into the estimated images by restricting information gains

from a pair of rays with slightly different viewpoints. The main idea of our algorithm is to make reconstructed scenes compact along individual rays and consistent across rays in the neighborhood. The proposed regularizers can be plugged into most of existing neural volume rendering techniques based on NeRF in a straightforward way. Despite its simplicity, we achieve consistently improved performance compared to existing neural view synthesis methods by large margins on multiple standard benchmarks.

****

Learning the Degradation Distribution for Blind Image Super-Resolution

Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6063-6072

Synthetic high-resolution (HR) & low-resolution (LR) pairs are widely used in existing super-resolution (SR) methods. To avoid the domain gap between synthetic and test images, most previous methods try to adaptively learn the synthesizing (degrading) process via a deterministic model. However, some degradations in real scenarios are stochastic and cannot be determined by the content of the image. These deterministic models may fail to model the random factors and content-independent parts of degradations, which will limit the performance of the following SR models. In this paper, we propose a probabilistic degradation model (PDM), which studies the degradation D as a random variable, and learns its distribution by modeling the mapping from a priori random variable z to D. Compared with previous deterministic degradation models, PDM could model more diverse degradations and generate HR-LR pairs that may better cover the various degradations of test images, and thus prevent the SR model from over-fitting to specific ones. Extensive experiments have demonstrated that our degradation model can help the SR model achieve better performance on different datasets. The source codes are released at git@github.com:greatlog/UnpairedSR.git.

****

Dist-PU: Positive-Unlabeled Learning From a Label Distribution Perspective

Yunrui Zhao, Qianqian Xu, Yangbangyan Jiang, Peisong Wen, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14461-14470

Positive-Unlabeled (PU) learning tries to learn binary classifiers from a few labeled positive examples with many unlabeled ones. Compared with ordinary semi-supervised learning, this task is much more challenging due to the absence of any known negative labels. While existing cost-sensitive-based methods have achieved state-of-the-art performances, they explicitly minimize the risk of classifying unlabeled data as negative samples, which might result in a negative-prediction preference of the classifier. To alleviate this issue, we resort to a label distribution perspective for PU learning in this paper. Noticing that the label distribution of unlabeled data is fixed when the class prior is known, it can be naturally used as supervision for the model. Motivated by this, we propose to pursue the label distribution consistency between predicted and ground-truth label distributions, which is formulated by aligning their expectations. Moreover, we further adopt the entropy minimization and Mixup regularization to avoid the trivial solution of the label distribution consistency on unlabeled data and mitigate the consequent confirmation bias. Experiments on three benchmark datasets validate the effectiveness of the proposed method.

****

SC2-PCR: A Second Order Spatial Compatibility for Efficient and Robust Point Cloud Registration

Zhi Chen, Kun Sun, Fan Yang, Wenbing Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13221-13231

In this paper, we present a second order spatial compatibility (SC^2) measure based method for efficient and robust point cloud registration (PCR), called SC^2-PCR. Firstly, we propose a second order spatial compatibility (SC^2) measure to compute the similarity between correspondences. It considers the global compatibility instead of local consistency, allowing for more distinctive clustering between inliers and outliers at early stage. Based on this measure, our registratio

n pipeline employs a global spectral technique to find some reliable seeds from the initial correspondences. Then we design a two-stage strategy to expand each seed to a consensus set based on the SC^2 measure matrix. Finally, we feed each consensus set to a weighted SVD algorithm to generate a candidate rigid transformation and select the best model as the final result. Our method can guarantee to find a certain number of outlier-free consensus sets using fewer samplings, making the model estimation more efficient and robust. In addition, the proposed SC^2 measure is general and can be easily plugged into deep learning based frameworks. Extensive experiments are carried out to investigate the performance of our method.

****

Relative Pose From a Calibrated and an Uncalibrated Smartphone Image
Yaqing Ding, Daniel Barath, Jian Yang, Zuzana Kukelova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12766-12775

In this paper, we propose a new minimal and a non-minimal solver for estimating the relative camera pose together with the unknown focal length of the second camera. This configuration has a number of practical benefits, e.g., when processing large-scale datasets. Moreover, it is resistant to the typical degenerate cases of the traditional six-point algorithm. The minimal solver requires four point correspondences and exploits the gravity direction that the built-in IMU of recent smart devices recover. We also propose a linear solver that enables estimating the pose from a larger-than-minimal sample extremely efficiently which then can be improved by, e.g., bundle adjustment. The methods are tested on 35654 image pairs from publicly available real-world datasets and the authors collected datasets. When combined with a recent robust estimator, they lead to results superior to the traditional solvers in terms of rotation, translation and focal length accuracy, while being notably faster.

****

Towards Robust and Reproducible Active Learning Using Neural Networks
Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, Shadab Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 223-232

Active learning (AL) is a promising ML paradigm that has the potential to parse through large unlabeled data and help reduce annotation cost in domains where labeling entire data can be prohibitive. Recently proposed neural network based AL methods use different heuristics to accomplish this goal. In this study, we demonstrate that under identical experimental conditions, different types of AL algorithms (uncertainty based, diversity based, and committee based) produce an inconsistent gain over random sampling baseline. Through a variety of experiments, controlling for sources of stochasticity, we show that variance in performance metrics achieved by AL algorithms can lead to results that are not consistent with the previously published results. We also found that under strong regularization, AL methods evaluated led to marginal or no advantage over the random sampling baseline under a variety of experimental conditions. Finally, we conclude with a set of recommendations on how to assess the results using a new AL algorithm to ensure results are reproducible and robust under changes in experimental conditions. We share our codes to facilitate AL experimentation. We believe our findings and recommendations will help advance reproducible research in AL using neural networks.

****

Retrieval Augmented Classification for Long-Tail Visual Recognition
Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6959-6969

We introduce Retrieval Augmented Classification (RAC), a generic approach to augmenting standard image classification pipelines with an explicit retrieval module. RAC consists of a standard base image encoder fused with a parallel retrieval branch that queries a non-parametric external memory of pre-encoded images and

associated text snippets. We apply RAC to the problem of long-tail classificatio n and demonstrate a significant improvement over previous state-of-the-art on Pl aces365-LT and iNaturalist-2018 (14.5% and 6.7% respectively), despite using onl y the training datasets themselves as the external information source. We demons trate that RAC's retrieval module, without prompting, learns a high level of acc uracy on tail classes. This, in turn, frees the base encoder to focus on common classes, and improve its performance thereon. RAC represents an alternative appr oach to utilizing large, pretrained models without requiring fine-tuning, as wel l as a first step towards more effectively making use of external memory within common computer vision architectures.
****

Not All Tokens Are Equal: Human-Centric Visual Analysis via Token Clustering Tra nsformer
Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, Xiaogang Wa ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 11101-11111
Vision transformers have achieved great successes in many computer vision tasks. Most methods generate vision tokens by splitting an image into a regular and fi xed grid and treating each cell as a token. However, not all regions are equally important in human-centric vision tasks, e.g., the human body needs a fine repr esentation with many tokens, while the image background can be modeled by a few tokens. To address this problem, we propose a novel Vision Transformer, called T oken Clustering Transformer (TCFormer), which merges tokens by progressive clust ering, where the tokens can be merged from different locations with flexible sha pes and sizes. The tokens in TCFormer can not only focus on important areas but also adjust the token shapes to fit the semantic concept and adopt a fine resolu tion for regions containing critical details, which is beneficial to capturing d etailed information. Extensive experiments show that TCFormer consistently outpe rforms its counterparts on different challenging humancentric tasks and datasets , including whole-body pose estimation on COCO-WholeBody and 3D human mesh recon struction on 3DPW. Code is available at https://github.com/ zengwang430521/TCFor mer.git.
****

Temporally Efficient Vision Transformer for Video Instance Segmentation
Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zha o, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2022, pp. 2885-2895
Recently vision transformer has achieved tremendous success on image-level visua l recognition tasks. To effectively and efficiently model the crucial temporal i nformation within a video clip, we propose a Temporally Efficient Vision Transfo rmer (TeViT) for video instance segmentation (VIS). Different from previous tran sformer-based VIS methods, TeViT is nearly convolution-free, which contains a tr ansformer backbone and a query-based video instance segmentation head. In the ba ckbone stage, we propose a nearly parameter-free messenger shift mechanism for e arly temporal context fusion. In the head stages, we propose a parameter-shared spatiotemporal query interaction mechanism to build the one-to-one correspondenc e between video instances and queries. Thus, TeViT fully utilizes both frame-lev el and instance-level temporal context information and obtains strong temporal m odeling capacity with negligible extra computational cost. On three widely adopt ed VIS benchmarks, i.e., YouTube-VIS-2019, YouTube-VIS-2021, and OVIS, TeViT obt ains state-of-the-art results and maintains high inference speed, e.g., 46.6 AP with 68.9 FPS on YouTube-VIS-2019. Code is available at https:// github.com/hust vl/TeViT.
****

The Devil Is in the Margin: Margin-Based Label Smoothing for Network Calibration
Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, Jose Dolz; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 80 -88
In spite of the dominant performances of deep neural networks, recent works have shown that they are poorly calibrated, resulting in over-confident predictions.

Miscalibration can be exacerbated by overfitting due to the minimization of the cross-entropy during training, as it promotes the predicted softmax probabilities to match the one-hot label assignments. This yields a pre-softmax activation of the correct class that is significantly larger than the remaining activations. Recent evidence from the literature suggests that loss functions that embed implicit or explicit maximization of the entropy of predictions yield state-of-the-art calibration performances. We provide a unifying constrained-optimization perspective of current state-of-the-art calibration losses. Specifically, these losses could be viewed as approximations of a linear penalty (or a Lagrangian term) imposing equality constraints on logit distances. This points to an important limitation of such underlying equality constraints, whose ensuing gradients constantly push towards a non-informative solution, which might prevent from reaching the best compromise between the discriminative performance and calibration of the model during gradient-based optimization. Following our observations, we propose a simple and flexible generalization based on inequality constraints, which imposes a controllable margin on logit distances. Comprehensive experiments on a variety of image classification, semantic segmentation and NLP benchmarks demonstrate that our method sets novel state-of-the-art results on these tasks in terms of network calibration, without affecting the discriminative performance. The code is available at https://github.com/by-liu/MbLS.

****

NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks

Fawaz Sammani, Tanmoy Mukherjee, Nikos Deligiannis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8322-8332

Natural language explanation (NLE) models aim at explaining the decision-making process of a black box system via generating natural language sentences which are human-friendly, high-level and fine-grained. Current NLE models explain the decision-making process of a vision or vision-language model (a.k.a., task model), e.g., a VQA model, via a language model (a.k.a., explanation model), e.g., GPT. Other than the additional memory resources and inference time required by the task model, the task and explanation models are completely independent, which dissociates the explanation from the reasoning process made to predict the answer. We introduce NLX-GPT, a general, compact and faithful language model that can simultaneously predict an answer and explain it. We first conduct pre-training on large scale data of image-caption pairs for general understanding of images, and then formulate the answer as a text prediction task along with the explanation. Without region proposals nor a task model, our resulting overall framework attains better evaluation scores, contains much less parameters and is 15x faster than the current SoA model. We then address the problem of evaluating the explanations which can be in many times generic, data-biased and can come in several forms. We therefore design 2 new evaluation measures: (1) explain-predict and (2) retrieval-based attack, a self-evaluation framework that requires no labels. Code is at: https://github.com/cvpr2022annonymous/nlxgpt

****

Bringing Old Films Back to Life

Ziyu Wan, Bo Zhang, Dongdong Chen, Jing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17694-17703

We present a learning-based framework, recurrent transformer network (RTN), to restore heavily degraded old films. Instead of performing frame-wise restoration, our method is based on the hidden knowledge learned from adjacent frames that contain abundant information about the occlusion, which is beneficial to restore challenging artifacts of each frame while ensuring temporal coherency. Moreover, contrasting the representation of the current frame and the hidden knowledge makes it possible to infer the scratch position in an unsupervised manner, and such defect localization generalizes well to real-world degradations. To better resolve mixed degradation and compensate for the flow estimation error during frame alignment, we propose to leverage more expressive transformer blocks for spatial restoration. Experiments on both synthetic dataset and real-world old films de

monstrate the significant superiority of the proposed RTN over existing solutions. In addition, the same framework can effectively propagate the color from keyframes to the whole video, ultimately yielding compelling restored films.
****

Sound and Visual Representation Learning With Multiple Pretraining Tasks
Arun Balajee Vasudevan, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14616-14626
Different self-supervised tasks (SSL) reveal different features from the data. The learned feature representations can exhibit different performance for each downstream task. In this light, this work aims to combine Multiple SSL tasks (Multi-SSL) that generalizes well for all downstream tasks. For this study, we investigate binaural sounds and image data. For binaural sounds, we propose three SSL tasks namely, spatial alignment, temporal synchronization of foreground objects and binaural audio and temporal gap prediction. We investigate several approaches of Multi-SSL and give insights into the downstream task performance on video retrieval, spatial sound super resolution, and semantic prediction on the OmniAudio dataset. Our experiments on binaural sound representations demonstrate that Multi-SSL via incremental learning (IL) of SSL tasks outperforms single SSL task models and fully supervised models in the downstream task performance. As a check of applicability on other modality, we also formulate our Multi-SSL models for image representation learning and we use the recently proposed SSL tasks, MoCov2 and DenseCL. Here, Multi-SSL surpasses recent methods such as MoCov2, DenseCL and DetCo by 2.06%, 3.27% and 1.19% on VOC07 classification and +2.83, +1.56 and +1.61 AP on COCO detection.
****

WarpingGAN: Warping Multiple Uniform Priors for Adversarial 3D Point Cloud Generation
Yingzhi Tang, Yue Qian, Qijian Zhang, Yiming Zeng, Junhui Hou, Xuefei Zhe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6397-6405
We propose WarpingGAN, an effective and efficient 3D point cloud generation network. Unlike existing methods that generate point clouds by directly learning the mapping functions between latent codes and 3D shapes, WarpingGAN learns a unified local-warping function to warp multiple identical pre-defined priors (i.e., sets of points uniformly distributed on regular 3D grids) into 3D shapes driven by local structure-aware semantics. In addition, we also ingeniously utilize the principle of the discriminator and tailor a stitching loss to eliminate the gaps between different partitions of a generated shape corresponding to different priors for boosting quality. Owing to the novel generating mechanism, WarpingGAN, a single lightweight network after onetime training, is capable of efficiently generating uniformly distributed 3D point clouds with various resolutions. Extensive experimental results demonstrate the superiority of our WarpingGAN over state-of-the-art methods to a large extent in terms of quantitative metrics, visual quality, and efficiency. The source code is publicly available at https://github.com/yztang4/WarpingGAN.git.
****

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models
Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11461-11471
Free-form inpainting is the task of adding new content to an image in the regions specified by an arbitrary binary mask. Most existing approaches train for a certain distribution of masks, which limits their generalization capabilities to unseen mask types. Furthermore, training with pixel-wise and perceptual losses often leads to simple textural extensions towards the missing areas instead of semantically meaningful generation. In this work, we propose RePaint: A Denoising Diffusion Probabilistic Model (DDPM) based inpainting approach that is applicable to even extreme masks. We employ a pretrained unconditional DDPM as the generative prior. To condition the generation process, we only alter the reverse diffus

ion iterations by sampling the unmasked regions using the given image informatio n. Since this technique does not modify or condition the original DDPM network i tself, the model produces high-quality and diverse output images for any inpaint ing form. We validate our method for both faces and general-purpose image inpain ting using standard and extreme masks. RePaint outperforms state-of-the-art Auto regressive, and GAN approaches for at least five out of six mask distributions. Github Repository: git.io/RePaint
****

Revealing Occlusions With 4D Neural Fields
Basile Van Hoorick, Purva Tendulkar, Dídac Surís, Dennis Park, Simon Stent, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3011-3021
For computer vision systems to operate in dynamic situations, they need to be ab le to represent and reason about object permanence. We introduce a framework for learning to estimate 4D visual representations from monocular RGB-D video, whic h is able to persist objects, even once they become obstructed by occlusions. Un like traditional video representations, we encode point clouds into a continuous representation, which permits the model to attend across the spatiotemporal con text to resolve occlusions. On two large video datasets that we release along wi th this paper, our experiments show that the representation is able to successfu lly reveal occlusions for several tasks, without any architectural changes. Visu alizations show that the attention mechanism automatically learns to follow occl uded objects. Since our approach can be trained end-to-end and is easily adaptab le, we believe it will be useful for handling occlusions in many video understan ding tasks. Data, code, and models are available at occlusions.cs.columbia.edu.
****

Meta Agent Teaming Active Learning for Pose Estimation
Jia Gong, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11079-11089
The existing pose estimation approaches often require a large number of annotate d images to attain good estimation performance, which are laborious to acquire. To reduce the human efforts on pose annotations, we propose a novel Meta Agent T eaming Active Learning (MATAL) framework to actively select and label informativ e images for effective learning. Our MATAL formulates the image selection proced ure as a Markov Decision Process and learns an optimal sampling policy that dire ctly maximizes the performance of the pose estimator. Our framework consists of a novel state-action representation as well as a multi-agent team to enable batc h sampling in the active learning procedure. The framework could be effectively optimized via Meta-Optimization to accelerate the adaptation to the gradually ex panded labeled data during deployment. Finally, we show experimental results on both human hand and body pose estimation benchmark datasets and demonstrate that our method significantly outperforms all baselines continuously under the same amount of annotation budget. Moreover, to obtain similar pose estimation accurac y, our MATAL framework can save around 40% labeling efforts on average compared to state-of-the-art active learning frameworks.
****

Forward Propagation, Backward Regression, and Pose Association for Hand Tracking in the Wild
Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, Minh Hoai; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6406-6416
We propose HandLer, a novel convolutional architecture that can jointly detect a nd track hands online in unconstrained videos. HandLer is based on Cascade-RCNNw ith additional three novel stages. The first stage is Forward Propagation, where the features from frame t-1 are propagated to frame t based on previously detec ted hands and their estimated motion. The second stage is the Detection and Back ward Regression, which uses outputs from the forward propagation to detect hands for frame t and their relative offset in frame t-1. The third stage uses an off -the-shelf human pose method to link any fragmented hand tracklets. We train the

forward propagation and backward regression and detection stages end-to-end together with the other Cascade-RCNN components.To train and evaluate HandLer, we also contribute YouTube-Hand, the first challenging large-scale dataset of unconstrained videos annotated with hand locations and their trajectories. Experiments on this dataset and other benchmarks show that HandLer outperforms the existing state-of-the-art tracking algorithms by a large margin.

****

Pseudo-Q: Generating Pseudo Language Queries for Visual Grounding
Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15513-15523

Visual grounding, i.e., localizing objects in images according to natural language queries, is an important topic in visual language understanding. The most effective approaches for this task are based on deep learning, which generally require expensive manually labeled image-query or patch-query pairs. To eliminate the heavy dependence on human annotations, we present a novel method, named Pseudo-Q, to automatically generate pseudo language queries for supervised training. Our method leverages an off-the-shelf object detector to identify visual objects from unlabeled images, and then language queries for these objects are obtained in an unsupervised fashion with a pseudo-query generation module. Then, we design a task-related query prompt module to specifically tailor generated pseudo language queries for visual grounding tasks. Further, in order to fully capture the contextual relationships between images and language queries, we develop a visual-language model equipped with multi-level cross-modality attention mechanism. Extensive experimental results demonstrate that our method has two notable benefits: (1) it can reduce human annotation costs significantly, e.g., 31% on RefCOCO without degrading original model's performance under the fully supervised setting, and (2) without bells and whistles, it achieves superior or comparable performance compared to state-of-the-art weakly-supervised visual grounding methods on all the five datasets we have experimented. Code is available at https://github.com/LeapLabTHU/Pseudo-Q.

****

E2(GO)MOTION: Motion Augmented Event Stream for Egocentric Action Recognition
Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, Barbara Caputo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19935-19947

Event cameras are novel bio-inspired sensors, which asynchronously capture pixel-level intensity changes in the form of "events". Due to their sensing mechanism, event cameras have little to no motion blur, a very high temporal resolution and require significantly less power and memory than traditional frame-based cameras. These characteristics make them a perfect fit to several real-world applications such as egocentric action recognition on wearable devices, where fast camera motion and limited power challenge traditional vision sensors. However, the ever-growing field of event-based vision has, to date, overlooked the potential of event cameras in such applications. In this paper, we show that event data is a very valuable modality for egocentric action recognition. To do so, we introduce N-EPIC-Kitchens, the first event-based camera extension of the large-scale EPIC-Kitchens dataset. In this context, we propose two strategies: (i) directly processing event-camera data with traditional video-processing architectures (E^2(GO)) and (ii) using event-data to distill optical flow information E^2(GO)MO). On our proposed benchmark, we show that event data provides a comparable performance to RGB and optical flow, yet without any additional flow computation at deploy time, and an improved performance of up to 4% with respect to RGB only information. The N-EPIC-Kitchens dataset is available at https://github.com/EgocentricVision/N-EPIC-Kitchens.

****

ES6D: A Computation Efficient and Symmetry-Aware 6D Pose Regression Framework
Ningkai Mo, Wanshui Gan, Naoto Yokoya, Shifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6718-6727

In this paper, a computation efficient regression framework is presented for est imating the 6D pose of rigid objects from a single RGB-D image, which is applica ble to handling symmetric objects. This framework is designed in a simple archit ecture that efficiently extracts point-wise features from RGB-D data using a ful ly convolutional network, called XYZNet, and directly regresses the 6D pose with out any post refinement. In the case of symmetric object, one object has multipl e ground-truth poses, and this one-to-many relationship may lead to estimation a mbiguity. In order to solve this ambiguity problem, we design a symmetry-invaria nt pose distance metric, called average (maximum) grouped primitives distance or A(M)GPD. The proposed A(M)GPD loss can make the regression network converge to the correct state, i.e., all minima in the A(M)GPD loss surface are mapped to th e correct poses. Extensive experiments on YCB-Video and T-LESS datasets demonstr ate the proposed framework's substantially superior performance in top accuracy and low computational cost.

****

Self-Supervised Deep Image Restoration via Adaptive Stochastic Gradient Langevin Dynamics

Weixi Wang, Ji Li, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2022, pp. 1989-1998

While supervised deep learning has been a prominent tool for solving many image restoration problems, there is an increasing interest on studying self-supervise d or un- supervised methods to address the challenges and costs of collecting tr uth images. Based on the neuralization of a Bayesian estimator of the problem, t his paper presents a self-supervised deep learning approach to general image res toration problems. The key ingredient of the neuralized estimator is an adaptive stochastic gradient Langevin dy- namics algorithm for efficiently sampling the posterior distri- bution of network weights. The proposed method is applied on t wo image restoration problems: compressed sensing and phase retrieval. The exper iments on these applications showed that the proposed method not only outperform ed existing non-learning and unsupervised solutions in terms of image restoratio n quality, but also is more computationally efficient.

****

Towards Discovering the Effectiveness of Moderately Confident Samples for Semi-S upervised Learning

Hui Tang, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14658-14667

Semi-supervised learning (SSL) has been studied for a long time to solve vision tasks in data-efficient application scenarios. SSL aims to learn a good classifi cation model using a few labeled data together with large-scale unlabeled data. Recent advances achieve the goal by combining multiple SSL techniques, e.g., sel f-training and consistency regularization. From unlabeled samples, they usually adopt a confidence filter (CF) to select reliable ones with high prediction conf idence. In this work, we study whether the moderately confident samples are usel ess and how to select the useful ones to improve model optimization. To answer t hese problems, we propose a novel Taylor expansion inspired filtration (TEIF) fr amework, which admits the samples of moderate confidence with similar feature or gradient to the respective one averaged over the labeled and highly confident u nlabeled data. It can produce a stable and new information induced network updat e, leading to better generalization. Two novel filters are derived from this fra mework and can be naturally explained in two perspectives. One is gradient synch ronization filter (GSF), which strengthens the optimization dynamic of fully-sup ervised learning; it selects the samples whose gradients are similar to class-wi se majority gradients. The other is prototype proximity filter (PPF), which invo lves more prototypical samples in training to learn better semantic representati ons; it selects the samples near class-wise prototypes. They can be integrated i nto SSL methods with CF. We use the state-of-the-art FixMatch as the baseline. E xperiments on popular SSL benchmarks show that we achieve the new state of the a rt.

****

OoD-Bench: Quantifying and Understanding Two Dimensions of Out-of-Distribution G

eneralization
Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zheng uo Li, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2022, pp. 7947-7958
Deep learning has achieved tremendous success with independent and identically d istributed (i.i.d.) data. However, the performance of neural networks often dege nerates drastically when encountering out-of-distribution (OoD) data, i.e., when training and test data are sampled from different distributions. While a pletho ra of algorithms have been proposed for OoD generalization, our understanding of the data used to train and evaluate these algorithms remains stagnant. In this work, we first identify and measure two distinct kinds of distribution shifts th at are ubiquitous in various datasets. Next, through extensive experiments, we c ompare OoD generalization algorithms across two groups of benchmarks, each domin ated by one of the distribution shifts, revealing their strengths on one shift a s well as limitations on the other shift. Overall, we position existing datasets and algorithms from different research areas seemingly unconnected into the sam e coherent picture. It may serve as a foothold that can be resorted to by future OoD generalization research. Our code is available at https://github.com/ynysjt u/ood_bench.
****

An Empirical Study of Training End-to-End Vision-and-Language Transformers
Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Cheng uang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, Michael Zeng; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 18166-18176
Vision-and-language (VL) pre-training has proven to be highly effective on vario us VL downstream tasks. While recent work has shown that fully transformer-based VL models can be more efficient than previous region-feature-based methods, the ir performance on downstream tasks often degrades significantly. In this paper, we present METER, a Multimodal End-to-end TransformER framework, through which w e investigate how to design and pre-train a fully transformer-based VL model in an end-to-end manner. Specifically, we dissect the model designs along multiple dimensions: vision encoders (e.g., CLIP-ViT, Swin transformer), text encoders (e .g., RoBERTa, DeBERTa), multimodal fusion module (e.g., merged attention vs. co- attention), architectural design (e.g., encoder-only vs. encoder-decoder), and p re-training objectives (e.g., masked image modeling). We conduct comprehensive e xperiments and provide insights on how to train a performant VL transformer whil e maintaining fast inference speed. Notably, our best model achieves an accuracy of 77.64% on the VQAv2 test-std set using only 4M images for pre-training, surp assing the state-of-the-art region-feature-based model by 1.04%, and outperformi ng the previous best fully transformer-based model by 1.6%.
****

Multimodal Dynamics: Dynamical Fusion for Trustworthy Multimodal Classification
Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, Jianhua Yao; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 20707-20717
Integration of heterogeneous and high-dimensional data (e.g., multiomics) is bec oming increasingly important. Existing multimodal classification algorithms main ly focus on improving performance by exploiting the complementarity from differe nt modalities. However, conventional approaches are basically weak in providing trustworthy multimodal fusion, especially for safety-critical applications (e.g. , medical diagnosis). For this issue, we propose a novel trustworthy multimodal classification algorithm termed Multimodal Dynamics, which dynamically evaluates both the feature-level and modality-level informativeness for different samples and thus trustworthily integrates multiple modalities. Specifically, a sparse g ating is introduced to capture the information variation of each within-modality feature and the true class probability is employed to assess the classification confidence of each modality. Then a transparent fusion algorithm based on the d ynamical informativeness estimation strategy is induced. To the best of our know ledge, this is the first work to jointly model both feature and modality variati

on for different samples to provide trustworthy fusion in multi-modal classification. Extensive experiments are conducted on multimodal medical classification datasets. In these experiments, superior performance and trustworthiness of our algorithm are clearly validated compared to the state-of-the-art methods.
****

The Neurally-Guided Shape Parser: Grammar-Based Labeling of 3D Shape Regions With Approximate Inference
R. Kenny Jones, Aalia Habib, Rana Hanocka, Daniel Ritchie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11614-11623
We propose the Neurally-Guided Shape Parser (NGSP), a method that learns how to assign fine-grained semantic labels to regions of a 3D shape. NGSP solves this problem via MAP inference, modeling the posterior probability of a label assignment conditioned on an input shape with a learned likelihood function. To make this search tractable, NGSP employs a neural guide network that learns to approximate the posterior. NGSP finds high-probability label assignments by first sampling proposals with the guide network and then evaluating each proposal under the full likelihood. We evaluate NGSP on the task of fine-grained semantic segmentation of manufactured 3D shapes from PartNet, where shapes have been decomposed into regions that correspond to part instance over-segmentations. We find that NGSP delivers significant performance improvements over comparison methods that (i) use regions to group per-point predictions, (ii) use regions as a self-supervisory signal or (iii) assign labels to regions under alternative formulations. Further, we show that NGSP maintains strong performance even with limited labeled data or noisy input shape regions. Finally, we demonstrate that NGSP can be directly applied to CAD shapes found in online repositories and validate its effectiveness with a perceptual study.
****

Unsupervised Homography Estimation With Coplanarity-Aware GAN
Mingbo Hong, Yuhang Lu, Nianjin Ye, Chunyu Lin, Qijun Zhao, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17663-17672
Estimating homography from an image pair is a fundamental problem in image alignment. Unsupervised learning methods have received increasing attention in this field due to their promising performance and label-free training. However, existing methods do not explicitly consider the problem of plane induced parallax, which will make the predicted homography compromised on multiple planes. In this work, we propose a novel method HomoGAN to guide unsupervised homography estimation to focus on the dominant plane. First, a multi-scale transformer network is designed to predict homography from the feature pyramids of input images in a coarse-to-fine fashion. Moreover, we propose an unsupervised GAN to impose coplanarity constraint on the predicted homography, which is realized by using a generator to predict a mask of aligned regions, and then a discriminator to check if two masked feature maps can be induced by a single homography. To validate the effectiveness of HomoGAN and its components, we conduct extensive experiments on a large-scale dataset, and results show that our matching error is 22% lower than the previous SOTA method. Our code will be publicly available.
****

LIFT: Learning 4D LiDAR Image Fusion Transformer for 3D Object Detection
Yihan Zeng, Da Zhang, Chunwei Wang, Zhenwei Miao, Ting Liu, Xin Zhan, Dayang Hao, Chao Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17172-17181
LiDAR and camera are two common sensors to collect data in time for 3D object detection under the autonomous driving context. Though the complementary information across sensors and time has great potential of benefiting 3D perception, taking full advantage of sequential cross-sensor data still remains challenging. In this paper, we propose a novel LiDAR Image Fusion Transformer (LIFT) to model the mutual interaction relationship of cross-sensor data over time. LIFT learns to align the input 4D sequential cross-sensor data to achieve multi-frame multi-modal information aggregation. To alleviate computational load, we project both po

int clouds and images into the bird-eye-view maps to compute sparse grid-wise se lf-attention. LIFT also benefits from a cross-sensor and cross-time data augment ation scheme. We evaluate the proposed approach on the challenging nuScenes and Waymo datasets, where our LIFT performs well over the state-of-the-art and stron g baselines.
****

AutoLoss-Zero: Searching Loss Functions From Scratch for Generic Tasks
Hao Li, Tianwen Fu, Jifeng Dai, Hongsheng Li, Gao Huang, Xizhou Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1009-1018
Significant progress has been achieved in automating the design of various compo nents in deep networks. However, the automatic design of loss functions for gene ric tasks with various evaluation metrics remains under-investigated. Previous w orks on handcrafting loss functions heavily rely on human expertise, which limit s their extensibility. Meanwhile, searching for loss functions is nontrivial due to the vast search space. Existing efforts mainly tackle the issue by employing task-specific heuristics on specific tasks and particular metrics. Such work ca nnot be extended to other tasks without arduous human effort. In this paper, we propose AutoLoss-Zero, which is a general framework for searching loss functions from scratch for generic tasks. Specifically, we design an elementary search sp ace composed only of primitive mathematical operators to accommodate the heterog eneous tasks and evaluation metrics. A variant of the evolutionary algorithm is employed to discover loss functions in the elementary search space. A loss-rejec tion protocol and a gradient-equivalence-check strategy are developed so as to i mprove the search efficiency, which are applicable to generic tasks. Extensive e xperiments on various computer vision tasks demonstrate that our searched loss f unctions are on par with or superior to existing loss functions, which generaliz e well to different datasets and networks. Code shall be released.
****

PatchNet: A Simple Face Anti-Spoofing Framework via Fine-Grained Patch Recogniti on
Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, Shang-Hong Lai; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20 281-20290
Face anti-spoofing (FAS) plays a critical role in securing face recognition syst ems from different presentation attacks. Previous works leverage auxiliary pixel -level supervision and domain generalization approaches to address unseen spoof types. However, the local characteristics of image captures, i.e., capturing dev ices and presenting materials, are ignored in existing works and we argue that s uch information is required for networks to discriminate between live and spoof images. In this work, we propose PatchNet which reformulates face anti-spoofing as a fine-grained patch-type recognition problem. To be specific, our framework recognizes the combination of capturing devices and presenting materials based o n the patches cropped from non-distorted face images. This reformulation can lar gely improve the data variation and enforce the network to learn discriminative feature from local capture patterns. In addition, to further improve the general ization ability of the spoof feature, we propose the novel Asymmetric Margin-bas ed Classification Loss and Self-supervised Similarity Loss to regularize the pat ch embedding space. Our experimental results verify our assumption and show that the model is capable of recognizing unseen spoof types robustly by only looking at local regions. Moreover, the fine-grained and patch-level reformulation of F AS outperforms the existing approaches on intra-dataset, cross-dataset, and doma in generalization benchmarks. Furthermore, our PatchNet framework can enable pra ctical applications like Few-Shot Reference-based FAS and facilitate future expl oration of spoof-related intrinsic cues.
****

OnePose: One-Shot Object Pose Estimation Without CAD Models
Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, X iaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 6825-6834

We propose a new method named OnePose for object pose estimation. Unlike existing instance-level or category-level methods, OnePose does not rely on CAD models and can handle objects in arbitrary categories without instance- or category-specific network training. OnePose draws the idea from visual localization and only requires a simple RGB video scan of the object to build a sparse SfM model of the object. Then, this model is registered to new query images with a generic feature matching network. To mitigate the slow runtime of existing visual localization methods, we propose a new graph attention network that directly matches 2D interest points in the query image with the 3D points in the SfM model, resulting in efficient and robust pose estimation. Combined with a feature-based pose tracker, OnePose is able to stably detect and track 6D poses of everyday household objects in real-time. We also collected a large-scale dataset that consists of 450 se- quences of 150 objects. Code and data are available at the project page: https://zju3dv.github.io/onepose/.

****

Weakly-Supervised Online Action Segmentation in Multi-View Instructional Videos
Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, Chiho Choi, Behzad Dariush; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13780-13790

This paper addresses a new problem of weakly-supervised online action segmentation in instructional videos. We present a framework to segment streaming videos online at test time using Dynamic Programming and show its advantages over greedy sliding window approach. We improve our framework by introducing the Online-Offline Discrepancy Loss (OODL) to encourage the segmentation results to have a higher temporal consistency. Furthermore, only during training, we exploit frame-wise correspondence between multiple views as supervision for training weakly-labeled instructional videos. In particular, we investigate three different multi-view inference techniques to generate more accurate frame-wise pseudo ground-truth with no additional annotation cost. We present results and ablation studies on two benchmark multi-view datasets, Breakfast and IKEA ASM. Experimental results show efficacy of the proposed methods both qualitatively and quantitatively in two domains of cooking and assembly.

****

Rethinking Minimal Sufficient Representation in Contrastive Learning
Haoqing Wang, Xun Guo, Zhi-Hong Deng, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16041-16050

Contrastive learning between different views of the data achieves outstanding success in the field of self-supervised representation learning and the learned representations are useful in broad downstream tasks. Since all supervision information for one view comes from the other view, contrastive learning approximately obtains the minimal sufficient representation which contains the shared information and eliminates the non-shared information between views. Considering the diversity of the downstream tasks, it cannot be guaranteed that all task-relevant information is shared between views. Therefore, we assume the non-shared task-relevant information cannot be ignored and theoretically prove that the minimal sufficient representation in contrastive learning is not sufficient for the downstream tasks, which causes performance degradation. This reveals a new problem that the contrastive learning models have the risk of over-fitting to the shared information between views. To alleviate this problem, we propose to increase the mutual information between the representation and input as regularization to approximately introduce more task-relevant information, since we cannot utilize any downstream task information during training. Extensive experiments verify the rationality of our analysis and the effectiveness of our method. It significantly improves the performance of several classic contrastive learning models in downstream tasks. Our code is available at https://github.com/Haoqing-Wang/InfoCL.

****

Disentangling Visual Embeddings for Attributes and Objects
Nirat Saini, Khoi Pham, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13658-13667

We study the problem of compositional zero-shot learning for object-attribute re

cognition. Prior works use visual features extracted with a backbone network, pre-trained for object classification and thus do not capture the subtly distinct features associated with attributes. To overcome this challenge, these studies employ supervision from the linguistic space, and use pre-trained word embeddings to better separate and compose attribute-object pairs for recognition. Analogous to linguistic embedding space, which already has unique and agnostic embeddings for object and attribute, we shift the focus back to the visual space and propose a novel architecture that can disentangle attribute and object features in the visual space. We use visual decomposed features to hallucinate embeddings that are representative for the seen and novel compositions to better regularize the learning of our model. Extensive experiments show that our method outperforms existing work with significant margin on three datasets: MIT-States, UT-Zappos, and a new benchmark created based on VAW.
****

Scalable Penalized Regression for Noise Detection in Learning With Noisy Labels
Yikai Wang, Xinwei Sun, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 346-355
Noisy training set usually leads to the degradation of generalization and robustness of neural networks. In this paper, we propose using a theoretically guaranteed noisy label detection framework to detect and remove noisy data for Learning with Noisy Labels (LNL). Specifically, we design a penalized regression to model the linear relation between network features and one-hot labels, where the noisy data are identified by the non-zero mean shift parameters solved in the regression model. To make the framework scalable to datasets that contain a large number of categories and training data, we propose a split algorithm to divide the whole training set into small pieces that can be solved by the penalized regression in parallel, leading to the Scalable Penalized Regression (SPR) framework. We provide the non-asymptotic probabilistic condition for SPR to correctly identify the noisy data. While SPR can be regarded as a sample selection module for standard supervised training pipeline, we further combine it with semi-supervised algorithm to further exploit the support of noisy data as unlabeled data. Experimental results on several benchmark datasets and real-world noisy datasets show the effectiveness of our framework. Our code and pretrained models are released at https://github.com/Yikai-Wang/SPR-LNL.
****

Effective Conditioned and Composed Image Retrieval Combining CLIP-Based Features
Alberto Baldrati, Marco Bertini, Tiberio Uricchio, Alberto Del Bimbo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21466-21474
Conditioned and composed image retrieval extend CBIR systems by combining a query image with an additional text that expresses the intent of the user, describing additional requests w.r.t. the visual content of the query image. This type of search is interesting for e-commerce applications, e.g. to develop interactive multimodal searches and chatbots. In this demo, we present an interactive system based on a combiner network, trained using contrastive learning, that combines visual and textual features obtained from the OpenAI CLIP network to address conditioned CBIR. The system can be used to improve e-shop search engines. For example, considering the fashion domain it lets users search for dresses, shirts and toptees using a candidate start image and expressing some visual differences w.r.t. its visual content, e.g. asking to change color, pattern or shape. The proposed network obtains state-of-the-art performance on the FashionIQ dataset and on the more recent CIRR dataset, showing its applicability to the fashion domain for conditioned retrieval, and to more generic content considering the more general task of composed image retrieval.
****

Registering Explicit to Implicit: Towards High-Fidelity Garment Mesh Reconstruction From Single Images
Heming Zhu, Lingteng Qiu, Yuda Qiu, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3845-3854
Fueled by the power of deep learning techniques and implicit shape learning, rec

ent advances in single-image human digitalization have reached unprecedented acc uracy and could recover fine-grained surface details such as garment wrinkles. H owever, a common problem for the implicit-based methods is that they cannot prod uce separated and topology-consistent mesh for each garment piece, which is cruc ial for the current 3D content creation pipeline. To address this issue, we prop osed a novel geometry inference framework ReEF that reconstructs topology- consi stent layered garment mesh by registering the explicit garment template to the w hole-body implicit fields predicted from single images. Experiments demonstrate that our method notably outperforms the counterparts on single-image layered gar ment reconstruction and could bring high-quality digital assets for further cont ent creation.

****

Federated Class-Incremental Learning

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, Qi Zhu; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 10164-10173

Federated learning (FL) has attracted growing attentions via data-private collab orative training on decentralized clients. However, most existing methods unreal istically assume object classes of the overall framework are fixed over time. It makes the global model suffer from significant catastrophic forgetting on old c lasses in real-world scenarios, where local clients often collect new classes co ntinuously and have very limited storage memory to store old classes. Moreover, new clients with unseen new classes may participate in the FL training, further aggravating the catastrophic forgetting of global model. To address these challe nges, we develop a novel Global-Local Forgetting Compensation (GLFC) model, to l earn a global class-incremental model for alleviating the catastrophic forgettin g from both local and global perspectives. Specifically, to address local forget ting caused by class imbalance at the local clients, we design a class-aware gra dient compensation loss and a class-semantic relation distillation loss to balan ce the forgetting of old classes and distill consistent inter-class relations ac ross tasks. To tackle the global forgetting brought by the non-i.i.d class imbal ance across clients, we propose a proxy server that selects the best old global model to assist the local relation distillation. Moreover, a prototype gradient-based communication mechanism is developed to protect the privacy. Our model out performs state-of-the-art methods by 4.4% 15.1% in terms of average accuracy on representative benchmark datasets. The code is available at https://github.com/c onditionWang/FCIL.

****

MiniViT: Compressing Vision Transformers With Weight Multiplexing

Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, Lu Yuan ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 12145-12154

Vision Transformer (ViT) models have recently drawn much attention in computer v ision due to their high model capability. However, ViT models suffer from huge n umber of parameters, restricting their applicability on devices with limited com putation. To alleviate this problem, we propose MiniViT, a new compression frame work, which achieves parameter reduction in vision transformers while retaining the same performance. The central idea of MiniViT is to multiplex the weights of consecutive transformer blocks. More specifically, we make the weights shared a cross layers, while imposing a transformation on the weights to increase diversi ty. Weight distillation over self-attention is also applied to transfer knowledg e from large-scale ViT models to weight-multiplexed compact models. Comprehensiv e experiments demonstrate the efficacy of MiniViT, showing that it can reduce th e size of the pre-trained Swin-B transformer by 48%, while achieving an increase of 1.0% in top-1 accuracy on ImageNet. Moreover, using a single-layer parameter s, MiniViT is able to compress DeiT-B by 9.7 times from 86M to 9M parameters, wi thout seriously compromising the performance. Finally, we verify the transferabi lity of MiniViT by reporting its performance on downstream benchmarks. Code and models are available at here.

****

Practical Stereo Matching via Cascaded Recurrent Network With Adaptive Correlation
Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16263-16272
With the advent of convolutional neural networks, stereo matching algorithms have recently gained tremendous progress. However, it remains a great challenge to accurately extract disparities from real-world image pairs taken by consumer-level devices like smartphones, due to practical complicating factors such as thin structures, non-ideal rectification, camera module inconsistencies and various hard-case scenes. In this paper, we propose a set of innovative designs to tackle the problem of practical stereo matching: 1) to better recover fine depth details, we design a hierarchical network with recurrent refinement to update disparities in a coarse-to-fine manner, as well as a stacked cascaded architecture for inference; 2) we propose an adaptive group correlation layer to mitigate the impact of erroneous rectification; 3) we introduce a new synthetic dataset with special attention to difficult cases for better generalizing to real-world scenes. Our results not only rank 1st on both Middlebury and ETH3D benchmarks, outperforming existing state-of-the-art methods by a notable margin, but also exhibit high-quality details for real-life photos, which clearly demonstrates the efficacy of our contributions.
****
D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions
Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20577-20586
We introduce the dynamic grasp synthesis task: given an object with a known 6D pose and a grasp reference, our goal is to generate motions that move the object to a target 6D pose. This is challenging, because it requires reasoning about the complex articulation of the human hand and the intricate physical interaction with the object. We propose a novel method that frames this problem in the reinforcement learning framework and leverages a physics simulation, both to learn and to evaluate such dynamic interactions. A hierarchical approach decomposes the task into low-level grasping and high-level motion synthesis. It can be used to generate novel hand sequences that approach, grasp, and move an object to a desired location, while retaining human-likeness. We show that our approach leads to stable grasps and generates a wide range of motions. Furthermore, even imperfect labels can be corrected by our method to generate dynamic interaction sequences.
****
Show, Deconfound and Tell: Image Captioning With Causal Inference
Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, Jiaqi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18041-18050
The transformer-based encoder-decoder framework has shown remarkable performance in image captioning. However, most transformer-based captioning methods ever overlook two kinds of elusive confounders: the visual confounder and the linguistic confounder, which generally lead to harmful bias, induce the spurious correlations during training, and degrade the model generalization. In this paper, we first use Structural Causal Models (SCMs) to show how two confounders damage the image captioning. Then we apply the backdoor adjustment to propose a novel causal inference based image captioning (CIIC) framework, which consists of an interventional object detector (IOD) and an interventional transformer decoder (ITD) to jointly confront both confounders. In the encoding stage, the IOD is able to disentangle the region-based visual features by deconfounding the visual confounder. In the decoding stage, the ITD introduces causal intervention into the transformer decoder and deconfounds the visual and linguistic confounders simultaneously. Two modules collaborate with each other to alleviate the spurious correlations caused by the unobserved confounders. When tested on MSCOCO, our proposal sig

nificantly outperforms the state-of-the-art encoder-decoder models on Karpathy s plit and online test split. Code is published in https: //github.com/CUMTGG/CIIC .

****

Extracting Triangular 3D Models, Materials, and Lighting From Images
Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Eva ns, Thomas Müller, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2022, pp. 8280-8290
We present an efficient method for joint optimization of topology, materials and lighting from multi-view image observations. Unlike recent multi-view reconstru ction approaches, which typically produce entangled 3D representations encoded i n neural networks, we output triangle meshes with spatially-varying materials an d environment lighting that can be deployed in any traditional graphics engine u nmodified. We leverage recent work in differentiable rendering, coordinate-based networks to compactly represent volumetric texturing, alongside differentiable marching tetrahedrons to enable gradient-based optimization directly on the surf ace mesh. Finally, we introduce a differentiable formulation of the split sum ap proximation of environment lighting to efficiently recover all-frequency lightin g. Experiments show our extracted models used in advanced scene editing, materia l decomposition, and high quality view interpolation, all running at interactive rates in triangle-based renderers (rasterizers and path tracers).

****

Weakly Supervised Segmentation on Outdoor 4D Point Clouds With Temporal Matching and Spatial Graph Propagation
Hanyu Shi, Jiacheng Wei, Ruibo Li, Fayao Liu, Guosheng Lin; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11840-11849
Existing point cloud segmentation methods require a large amount of annotated da ta, especially for the outdoor point cloud scene. Due to the complexity of the o utdoor 3D scenes, manual annotations on the outdoor point cloud scene are time-c onsuming and expensive. In this paper, we study how to achieve scene understandi ng with limited annotated data. Treating 100 consecutive frames as a sequence, w e divide the whole dataset into a series of sequences and annotate only 0.1% poi nts in the first frame of each sequence to reduce the annotation requirements. T his leads to a total annotation budget of 0.001%. We propose a novel temporal-sp atial framework for effective weakly supervised learning to generate high-qualit y pseudo labels from these limited annotated data. Specifically, the framework c ontains two modules: an matching module in temporal dimension to propagate pseud o labels across different frames, and a graph propagation module in spatial dime nsion to propagate the information of pseudo labels to the entire point clouds i n each frame. With only 0.001% annotations for training, experimental results on both SemanticKITTI and SemanticPOSS shows our weakly supervised two-stage frame work is comparable to some existing fully supervised methods. We also evaluate o ur framework with 0.005% initial annotations on SemanticKITTI, and achieve a res ult close to fully supervised backbone model.

****

ImFace: A Nonlinear 3D Morphable Face Model With Implicit Neural Representations
Mingwu Zheng, Hongyu Yang, Di Huang, Liming Chen; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20343-2035 2
Precise representations of 3D faces are beneficial to various computer vision an d graphics applications. Due to the data discretization and model linearity howe ver, it remains challenging to capture accurate identity and expression clues in current studies. This paper presents a novel 3D morphable face model, namely Im Face, to learn a nonlinear and continuous space with implicit neural representat ions. It builds two explicitly disentangled deformation fields to model complex shapes associated with identities and expressions, respectively, and designs an improved learning strategy to extend embeddings of expressions to allow more div erse changes. We further introduce a Neural Blend-Field to learn sophisticated d etails by adaptively blending a series of local fields. In addition to ImFace, a

n effective preprocessing pipeline is proposed to address the issue of watertigh t input requirement in implicit representations, enabling them to work with comm on facial surfaces for the first time. Extensive experiments are performed to de monstrate the superiority of ImFace.
****

MobRecon: Mobile-Friendly Hand Mesh Reconstruction From Monocular Image
Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Y uan Zhang, Xiaoyan Guo; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 20544-20554
In this work, we propose a framework for single-view hand mesh reconstruction, w hich can simultaneously achieve high reconstruction accuracy, fast inference spe ed, and temporal coherence. Specifically, for 2D encoding, we propose lightweigh t yet effective stacked structures. Regarding 3D decoding, we provide an efficie nt graph operator, namely depth-separable spiral convolution. Moreover, we prese nt a novel feature lifting module for bridging the gap between 2D and 3D represe ntations. This module begins with a map-based position regression (MapReg) block to integrate the merits of both heatmap encoding and position regression paradi gms for improved 2D accuracy and temporal coherence. Furthermore, MapReg is foll owed by pose pooling and pose-to-vertex lifting approaches, which transform 2D p ose encodings to semantic features of 3D vertices. Overall, our hand reconstruct ion framework, called MobRecon, comprises affordable computational costs and min iature model size, which reaches a high inference speed of 83FPS on Apple A14 CP U. Extensive experiments on popular datasets such as FreiHAND, RHD, and HO3Dv2 d emonstrate that our MobRecon achieves superior performance on reconstruction acc uracy and temporal coherence. Our code is publicly available at https://github.c om/SeanChenxy/HandMesh.
****

Layered Depth Refinement With Mask Guidance
Soo Ye Kim, Jianming Zhang, Simon Niklaus, Yifei Fan, Simon Chen, Zhe Lin, Munch url Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 3855-3865
Depth maps are used in a wide range of applications from 3D rendering to 2D imag e effects such as Bokeh. However, those predicted by single image depth estimati on (SIDE) models often fail to capture isolated holes in objects and/or have ina ccurate boundary regions. Meanwhile, high-quality masks are much easier to obtai n, using commercial auto-masking tools or off-the-shelf methods of segmentation and matting or even by manual editing. Hence, in this paper, we formulate a nove l problem of mask-guided depth refinement that utilizes a generic mask to refine the depth prediction of SIDE models. Our framework performs layered refinement and inpainting/outpainting, decomposing the depth map into two separate layers s ignified by the mask and the inverse mask. As datasets with both depth and mask annotations are scarce, we propose a self-supervised learning scheme that uses a rbitrary masks and RGB-D datasets. We empirically show that our method is robust to different types of masks and initial depth predictions, accurately refining depth values in inner and outer mask boundary regions. We further analyze our mo del with an ablation study and demonstrate results on real applications.
****

Parameter-Free Online Test-Time Adaptation
Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, Luca Bertinetto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8344-8353
Training state-of-the-art vision models has become prohibitively expensive for r esearchers and practitioners. For the sake of accessibility and resource reuse, it is important to focus on adapting these models to a variety of downstream sce narios. An interesting and practical paradigm is online test-time adaptation, ac cording to which training data is inaccessible, no labelled data from the test d istribution is available, and adaptation can only happen at test time and on a h andful of samples. In this paper, we investigate how test-time adaptation method s fare for a number of pre-trained models on a variety of real-world scenarios, significantly extending the way they have been originally evaluated. We show tha

t they perform well only in narrowly-defined experimental setups and sometimes f
ail catastrophically when their hyperparameters are not selected for the same sc
enario in which they are being tested. Motivated by the inherent uncertainty aro
und the conditions that will ultimately be encountered at test time, we propose
a particularly "conservative" approach, which addresses the problem with a Lapla
cian Adjusted Maximum-likelihood Estimation (LAME) objective. By adapting the mo
del's output (not its parameters), and solving our objective with an efficient c
oncave-convex procedure, our approach exhibits a much higher average accuracy ac
ross scenarios than existing methods, while being notably faster and have a much
 lower memory footprint. The code is available at https://github.com/fiveai/LAME
.
****

SIGMA: Semantic-Complete Graph Matching for Domain Adaptive Object Detection
Wuyang Li, Xinyu Liu, Yixuan Yuan; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2022, pp. 5291-5300
Domain Adaptive Object Detection (DAOD) leverages a labeled domain to learn an o
bject detector generalizing to a novel domain free of annotations. Recent advanc
es align class-conditional distributions by narrowing down cross-domain prototyp
es (class centers). Though great success, they ignore the significant within-cla
ss variance and the domain-mismatched semantics within the training batch, leadi
ng to a sub-optimal adaptation. To overcome these challenges, we propose a novel
 SemantIc-complete Graph MAtching (SIGMA) framework for DAOD, which completes mi
smatched semantics and reformulates the adaptation with graph matching. Specific
ally, we design a Graph-embedded Semantic Completion module (GSC) that completes
 mismatched semantics through generating hallucination graph nodes in missing ca
tegories. Then, we establish cross-image graphs to model class-conditional distr
ibutions and learn a graph-guided memory bank for better semantic completion in
turn. After representing the source and target data as graphs, we reformulate th
e adaptation as a graph matching problem, i.e., finding well-matched node pairs
across graphs to reduce the domain gap, which is solved with a novel Bipartite G
raph Matching adaptor (BGM). In a nutshell, we utilize graph nodes to establish
semantic-aware node affinity and leverage graph edges as quadratic constraints i
n a structure-aware matching loss, achieving fine-grained adaptation with a node
-to-node graph matching. Extensive experiments verify that SIGMA outperforms exi
sting works significantly. Our codes are available at https://github.com/CityU-A
IM-Group/SIGMA.
****

Global Convergence of MAML and Theory-Inspired Neural Architecture Search for Fe
w-Shot Learning
Haoxiang Wang, Yite Wang, Ruoyu Sun, Bo Li; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9797-9808
Model-agnostic meta-learning (MAML) and its variants have become popular approac
hes for few-shot learning. However, due to the non-convexity of deep neural nets
 (DNNs) and the bi-level formulation of MAML, the theoretical properties of MAML
 with DNNs remain largely unknown. In this paper, we first prove that MAML with
over-parameterized DNNs is guaranteed to converge to global optima at a linear r
ate. Our convergence analysis indicates that MAML with over-parameterized DNNs i
s equivalent to kernel regression with a novel class of kernels, which we name a
s Meta Neural Tangent Kernels (MetaNTK). Then, we propose MetaNTK-NAS, a new tra
ining-free neural architecture search (NAS) method for few-shot learning that us
es MetaNTK to rank and select architectures. Empirically, we compare our MetaNTK
-NAS with previous NAS methods on two popular few-shot learning benchmarks, mini
ImageNet, and tieredImageNet. We show that the performance of MetaNTK-NAS is com
parable or better than the state-of-the-art NAS method designed for few-shot lea
rning while enjoying more than 100x speedup. We believe the efficiency of MetaNT
K-NAS makes itself more practical for many real-world tasks.
****

LAKe-Net: Topology-Aware Point Cloud Completion by Localizing Aligned Keypoints
Junshu Tang, Zhijun Gong, Ran Yi, Yuan Xie, Lizhuang Ma; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 172

6-1735
Point cloud completion aims at completing geometric and topological shapes from a partial observation. However, some topology of the original shape is missing, existing methods directly predict the location of complete points, without predicting structured and topological information of the complete shape, which leads to inferior performance. To better tackle the missing topology part, we propose LAKe-Net, a novel topology-aware point cloud completion model by localizing aligned keypoints, with a novel Keypoints-Skeleton-Shape prediction manner. Specifically, our method completes missing topology using three steps: Aligned Keypoint Localization. An asymmetric keypoint locator, including an unsupervised multi-scale keypoint detector and a complete keypoint generator, is proposed for localizing aligned keypoints from complete and partial point clouds. We theoretically prove that the detector can capture aligned keypoints for objects within a sub-category. Surface-skeleton Generation. A new type of skeleton, named Surface-skeleton, is generated from keypoints based on geometric priors to fully represent the topological information captured from keypoints and better recover the local details. Shape Refinement. We design a refinement subnet where multi-scale surface-skeletons are fed into each recursive skeleton-assisted refinement module to assist the completion process. Experimental results show that our method achieves the state-of-the-art performance on point cloud completion.
****

Scribble-Supervised LiDAR Semantic Segmentation
Ozan Unal, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2697-2707
Densely annotating LiDAR point clouds remains too expensive and time-consuming to keep up with the ever growing volume of data. While current literature focuses on fully-supervised performance, developing efficient methods that take advantage of realistic weak supervision have yet to be explored. In this paper, we propose using scribbles to annotate LiDAR point clouds and release ScribbleKITTI, the first scribble-annotated dataset for LiDAR semantic segmentation. Furthermore, we present a pipeline to reduce the performance gap that arises when using such weak annotations. Our pipeline comprises of three stand-alone contributions that can be combined with any LiDAR semantic segmentation model to achieve up to 95.7% of the fully-supervised performance while using only 8% labeled points. Our scribble annotations and code are available at github.com/ouenal/scribblekitti.
****

AlignMixup: Improving Representations by Interpolating Aligned Features
Shashanka Venkataramanan, Ewa Kijak, Laurent Amsaleg, Yannis Avrithis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19174-19183
Mixup is a powerful data augmentation method that interpolates between two or more examples in the input or feature space and between the corresponding target labels. However, how to best interpolate images is not well defined. Recent mixup methods overlay or cut-and-paste two or more objects into one image, which needs care in selecting regions. Mixup has also been connected to autoencoders, because often autoencoders generate an image that continuously deforms into another. However, such images are typically of low quality. In this work, we revisit mixup from the deformation perspective and introduce AlignMixup, where we geometrically align two images in the feature space. The correspondences allow us to interpolate between two sets of features, while keeping the locations of one set. Interestingly, this retains mostly the geometry or pose of one image and the appearance or texture of the other. We also show that an autoencoder can still improve representation learning under mixup, without the classifier ever seeing decoded images. AlignMixup outperforms state-of-the-art mixup methods on five different benchmarks.
****

No Pain, Big Gain: Classify Dynamic Point Cloud Sequences With Static Models by Fitting Feature-Level Space-Time Surfaces
Jia-Xing Zhong, Kaichen Zhou, Qingyong Hu, Bing Wang, Niki Trigoni, Andrew Markham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn

ition (CVPR), 2022, pp. 8510-8520
Scene flow is a powerful tool for capturing the motion field of 3D point clouds. However, it is difficult to directly apply flow-based models to dynamic point cloud classification since the unstructured points make it hard or even impossible to efficiently and effectively trace point-wise correspondences. To capture 3D motions without explicitly tracking correspondences, we propose a kinematics-inspired neural network (Kinet) by generalizing the kinematic concept of ST-surfaces to the feature space. By unrolling the normal solver of ST-surfaces in the feature space, Kinet implicitly encodes feature-level dynamics and gains advantages from the use of mature backbones for static point cloud processing. With only minor changes in network structures and low computing overhead, it is painless to jointly train and deploy our framework with a given static model. Experiments on NvGesture, SHREC'17, MSRAction-3D, and NTU-RGBD demonstrate its efficacy in performance, efficiency in both the number of parameters and computational complexity, as well as its versatility to various static backbones. Noticeably, Kinet achieves the accuracy of 93.27% on MSRAction-3D with only 3.20M parameters and 10.35G FLOPS. The code is available at https://github.com/jx-zhong-for-academic-purpose/Kinet.
****

HiVT: Hierarchical Vector Transformer for Multi-Agent Motion Prediction
Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, Kejie Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8823-8833
Accurately predicting the future motions of surrounding traffic agents is critical for the safety of autonomous vehicles. Recently, vectorized approaches have dominated the motion prediction community due to their capability of capturing complex interactions in traffic scenes. However, existing methods neglect the symmetries of the problem and suffer from the expensive computational cost, facing the challenge of making real-time multi-agent motion prediction without sacrificing the prediction performance. To tackle this challenge, we propose Hierarchical Vector Transformer (HiVT) for fast and accurate multi-agent motion prediction. By decomposing the problem into local context extraction and global interaction modeling, our method can effectively and efficiently model a large number of agents in the scene. Meanwhile, we propose a translation-invariant scene representation and rotation-invariant spatial learning modules, which extract features robust to the geometric transformations of the scene and enable the model to make accurate predictions for multiple agents in a single forward pass. Experiments show that HiVT achieves the state-of-the-art performance on the Argoverse motion forecasting benchmark with a small model size and can make fast multi-agent motion prediction.
****

HerosNet: Hyperspectral Explicable Reconstruction and Optimal Sampling Deep Network for Snapshot Compressive Imaging
Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17532-17541
Hyperspectral imaging is an essential imaging modality for a wide range of applications, especially in remote sensing, agriculture, and medicine. Inspired by existing hyperspectral cameras that are either slow, expensive, or bulky, reconstructing hyperspectral images (HSIs) from a low-budget snapshot measurement has drawn wide attention. By mapping a truncated numerical optimization algorithm into a network with a fixed number of phases, recent deep unfolding networks (DUNs) for spectral snapshot compressive sensing (SCI) have achieved remarkable success. However, DUNs are far from reaching the scope of industrial applications limited by the lack of cross-phase feature interaction and adaptive parameter adjustment. In this paper, we propose a novel Hyperspectral Explicable Reconstruction and Optimal Sampling deep Network for SCI, dubbed HerosNet, which includes several phases under the ISTA-unfolding framework. Each phase can flexibly simulate the sensing matrix and contextually adjust the step size in the gradient descent step, and hierarchically fuse and interact the hidden states of previous phases t

o effectively recover current HSI frames in the proximal mapping step. Simultaneously, a hardware-friendly optimal binary mask is learned end-to-end to further improve the reconstruction performance. Finally, our HerosNet is validated to outperform the state-of-the-art methods on both simulation and real datasets by large margins. The source code is available at https://github.com/jianzhangcs/HerosNet.
****

Vision Transformer Slimming: Multi-Dimension Searching in Continuous Optimization Space
Arnav Chavan, Zhiqiang Shen, Zhuang Liu, Zechun Liu, Kwang-Ting Cheng, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4931-4941
This paper explores the feasibility of finding an optimal sub-model from a vision transformer and introduces a pure vision transformer slimming (ViT-Slim) framework. It can search a sub-structure from the original model end-to-end across multiple dimensions, including the input tokens, MHSA and MLP modules with state-of-the-art performance. Our method is based on a learnable and unified l1 sparsity constraint with pre-defined factors to reflect the global importance in the continuous searching space of different dimensions. The searching process is highly efficient through a single-shot training scheme. For instance, on DeiT-S, ViT-Slim only takes 43 GPU hours for the searching process, and the searched structure is flexible with diverse dimensionalities in different modules. Then, a budget threshold is employed according to the requirements of accuracy-FLOPs trade-off on running devices, and a re-training process is performed to obtain the final model. The extensive experiments show that our ViT-Slim can compress up to 40% of parameters and 40% FLOPs on various vision transformers while increasing the accuracy by 0.6% on ImageNet. We also demonstrate the advantage of our searched models on several downstream datasets. Our code is available at https://github.com/Arnav0400/ViT-Slim.
****

Brain-Inspired Multilayer Perceptron With Spiking Neurons
Wenshuo Li, Hanting Chen, Jianyuan Guo, Ziyang Zhang, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 783-793
Recently, Multilayer Perceptron (MLP) becomes the hotspot in the field of computer vision tasks. Without inductive bias, MLPs perform well on feature extraction and achieve amazing results. However, due to the simplicity of their structures, the performance highly depends on the local features communication machenism. To further improve the performance of MLP, we introduce information communication mechanisms from brain-inspired neural networks. Spiking Neural Network (SNN) is the most famous brain-inspired neural network, and achieve great success on dealing with sparse data. Leaky Integrate and Fire (LIF) neurons in SNNs are used to communicate between different time steps. In this paper, we incorporate the machanism of LIF neurons into the MLP models, to achieve better accuracy without extra FLOPs. We propose a full-precision LIF operation to communicate between patches, including horizontal LIF and vertical LIF in different directions. We also propose to use group LIF to extract better local features. With LIF modules, our SNN-MLP model achieves 81.9%, 83.3% and 83.5% top-1 accuracy on ImageNet dataset with only 4.4G, 8.5G and 15.2G FLOPs, respectively, which are state-of-the-art results as far as we know.
****

Learning To Estimate Robust 3D Human Mesh From In-the-Wild Crowded Scenes
Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1475-1484
We consider the problem of recovering a single person's 3D human mesh from in-the-wild crowded scenes. While much progress has been in 3D human mesh estimation, existing methods struggle when test input has crowded scenes. The first reason for the failure is a domain gap between training and testing data. A motion capture dataset, which provides accurate 3D labels for training, lacks crowd data an

d impedes a network from learning crowded scene-robust image features of a targe
t person. The second reason is a feature processing that spatially averages the
feature map of a localized bounding box containing multiple people. Averaging th
e whole feature map makes a target person's feature indistinguishable from other
s. We present 3DCrowdNet that firstly explicitly targets in-the-wild crowded sce
nes and estimates a robust 3D human mesh by addressing the above issues. First,
we leverage 2D human pose estimation that does not require a motion capture data
set with 3D labels for training and does not suffer from the domain gap. Second,
 we propose a joint-based regressor that distinguishes a target person's feature
 from others. Our joint-based regressor preserves the spatial activation of a ta
rget by sampling features from the target's joint locations and regresses human
model parameters. As a result, 3DCrowdNet learns target-focused features and eff
ectively excludes the irrelevant features of nearby persons. We conduct experime
nts on various benchmarks and prove the robustness of 3DCrowdNet to the in-the-w
ild crowded scenes both quantitatively and qualitatively. Codes are available he
re: https://github.com/hongsukchoi/3DCrowdNet_RELEASE
****

ObjectFormer for Image Manipulation Detection and Localization
Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam
Lim, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2022, pp. 2364-2373
Recent advances in image editing techniques have posed serious challenges to the
 trustworthiness of multimedia data, which drives the research of image tamperin
g detection. In this paper, we propose ObjectFormer to detect and localize image
 manipulations. To capture subtle manipulation traces that are no longer visible
 in the RGB domain, we extract the high-frequency features of the images and com
bine them with RGB features as multimodal patch embeddings. In order to detect f
orgery traces, we use a set of learnable object prototypes as mid-level represen
tations to model the object-level consistencies among different regions, which a
re further used to refine patch embeddings to capture the patch-level consistenc
ies. We conduct extensive experiments on various datasets and the results verify
 the effectiveness of the proposed method, outperforming state-of-the-art tamper
ing detection and localization methods.
****

Detecting Deepfakes With Self-Blended Images
Kaede Shiohara, Toshihiko Yamasaki; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 18720-18729
In this paper, we present novel synthetic training data called self-blended imag
es (SBIs) to detect deepfakes. SBIs are generated by blending pseudo source and
target images from single pristine images, reproducing common forgery artifacts
(e.g., blending boundaries and statistical inconsistencies between source and ta
rget images). The key idea behind SBIs is that more general and hardly recogniza
ble fake samples encourage classifiers to learn generic and robust representatio
ns without overfitting to manipulation-specific artifacts. We compare our approa
ch with state-of- the-art methods on FF++, CDF, DFD, DFDC, DFDCP, and FFIW datas
ets by following the standard cross-dataset and cross-manipulation protocols. Ex
tensive experiments show that our method improves the model generalization to un
known manipulations and scenes. In particular, on DFDC and DFDCP where existing
methods suffer from the domain gap between the training and test sets, our appro
ach outperforms the baseline by 4.90% and 11.78% points in the cross-dataset eva
luation, respectively. Code is available at https://github.com/mapooon/SelfBlend
edImages.
****

Correlation-Aware Deep Tracking
Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, Wenjun Zeng; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2022, pp. 8751-8760
Robustness and discrimination power are two fundamental requirements in visual o
bject tracking. In most tracking paradigms, we find that the features extracted
by the popular Siamese-like networks can not fully discriminatively model the tr

acked targets and distractor objects, hindering them from simultaneously meeting these two requirements. While most methods focus on designing robust matching operations, we propose a novel target-dependent feature network inspired by the self-/cross-attention scheme. In contrast to the Siamese-like feature extraction, our network deeply embeds cross-image feature correlation in multiple layers of the feature network. By extensively matching the features of the two images through multiple layers, it is able to suppress non-target features, resulting in instance-varying feature extraction. The output features of the search image can be directly used for predicting target locations without extra correlation step. Moreover, our model can be flexibly pre-trained on abundant unpaired images, leading to notably faster convergence than the existing methods. Extensive experiments show our method achieves the state-of-the-art results while running at real-time. Our feature networks also can be applied to existing tracking pipelines seamlessly to raise the tracking performance.

****

Learnable Irrelevant Modality Dropout for Multimodal Action Recognition on Modality-Specific Annotated Videos

Saghir Alfasly, Jian Lu, Chen Xu, Yuru Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20208-20217

With the assumption that a video dataset is multimodality annotated in which auditory and visual modalities both are labeled or class-relevant, current multimodal methods apply modality fusion or cross-modality attention. However, effectively leveraging the audio modality in vision-specific annotated videos for action recognition is of particular challenge. To tackle this challenge, we propose a novel audio-visual framework that effectively leverages the audio modality in any solely vision-specific annotated dataset. We adopt the language models (e.g., BERT) to build a semantic audio-video label dictionary (SAVLD) that maps each video label to its most K-relevant audio labels in which SAVLD serves as a bridge between audio and video datasets. Then, SAVLD along with a pretrained audio multi-label model are used to estimate the audio-visual modality relevance during the training phase. Accordingly, a novel learnable irrelevant modality dropout (IMD) is proposed to completely drop out the irrelevant audio modality and fuse only the relevant modalities. Moreover, we present a new two-stream video Transformer for efficiently modeling the visual modalities. Results on several vision-specific annotated datasets including Kinetics400 and UCF-101 validated our framework as it outperforms most relevant action recognition methods.

****

NeurMiPs: Neural Mixture of Planar Experts for View Synthesis

Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, Shenlong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15702-15712

We present Neural Mixtures of Planar Experts (NeurMiPs), a novel planar-based scene representation for modeling geometry and appearance. NeurMiPs leverages a collection of local planar experts in 3D space as the scene representation. Each planar expert consists of the parameters of the local rectangular shape representing geometry and a neural radiance field modeling the color and opacity. We render novel views by calculating ray-plane intersections and composite output colors and densities at intersected points to the image. NeurMiPs blends the efficiency of explicit mesh rendering and flexibility of the neural radiance field. Experiments demonstrate that our proposed method achieves superior performance and speedup compared to other 3D representations for novel view synthesis.

****

Implicit Sample Extension for Unsupervised Person Re-Identification

Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7369-7378

Most existing unsupervised person re-identification (Re-ID) methods use clustering to generate pseudo labels for model training. Unfortunately, clustering sometimes mixes different true identities together or splits the same identity into two or more sub clusters. Training on these noisy clusters substantially hampers

the Re-ID accuracy. Due to the limited samples in each identity, we suppose ther e may lack some underlying information to well reveal the accurate clusters. To discover these information, we propose an Implicit Sample Extension (ISE) method to generate what we call support samples around the cluster boundaries. Specifi cally, we generate support samples from actual samples and their neighbouring cl usters in the embedding space through a progressive linear interpolation (PLI) s trategy. PLI controls the generation with two critical factors, i.e., 1) the dir ection from the actual sample towards its K-nearest clusters and 2) the degree f or mixing up the context information from the K-nearest clusters. Meanwhile, giv en the support samples, ISE further uses a label-preserving loss to pull them to wards their corresponding actual samples, so as to compact each cluster. Consequ ently, ISE reduces the "sub and mixed" clustering errors, thus improving the Re-ID performance. Extensive experiments demonstrate that the proposed method is ef fective and achieves state-of-the-art performance for unsupervised person Re-ID. Code is available at: https://github.com/PaddlePaddle/PaddleClas.

****

Energy-Based Latent Aligner for Incremental Learning

K J Joseph, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, Vineeth N Balas ubramanian; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2022, pp. 7452-7461

Deep learning models tend to forget their earlier knowledge while incrementally learning new tasks. This behavior emerges because the parameter updates optimize d for the new tasks may not align well with the updates suitable for older tasks . The resulting latent representation mismatch causes forgetting. In this work, we propose ELI: Energy-based Latent Aligner for Incremental Learning, which firs t learns an energy manifold for the latent representations such that previous ta sk latents will have low energy and the current task latents have high energy va lues. This learned manifold is used to counter the representational shift that h appens during incremental learning. The implicit regularization that is offered by our proposed methodology can be used as a plug-and-play module in existing in cremental learning methodologies. We validate this through extensive evaluation on CIFAR-100, ImageNet subset, ImageNet 1k and Pascal VOC datasets. We observe c onsistent improvement when ELI is added to three prominent methodologies in clas s-incremental learning, across multiple incremental settings. Further, when adde d to the state-of-the-art incremental object detector, ELI provides over 5% impr ovement in detection accuracy, corroborating its effectiveness and complementary advantage to existing art. Code is available at: https://github.com/JosephKJ/EL I.

****

Towards Semi-Supervised Deep Facial Expression Recognition With an Adaptive Conf idence Margin

Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, Xinbo Gao; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 416 6-4175

Only parts of unlabeled data are selected to train models for most semi-supervis ed learning methods, whose confidence scores are usually higher than the pre-def ined threshold (i.e., the confidence margin). We argue that the recognition perf ormance should be further improved by making full use of all unlabeled data. In this paper, we learn an Adaptive Confidence Margin (Ada-CM) to fully leverage al l unlabeled data for semi-supervised deep facial expression recognition. All unl abeled samples are partitioned into two subsets by comparing their confidence sc ores with the adaptively learned confidence margin at each training epoch: (1) s ubset I including samples whose confidence scores are no lower than the margin; (2) subset II including samples whose confidence scores are lower than the margi n. For samples in subset I, we constrain their predictions to match pseudo label s. Meanwhile, samples in subset II participate in the feature-level contrastive objective to learn effective facial expression features. We extensively evaluate Ada-CM on four challenging datasets, showing that our method achieves state-of-the-art performance, especially surpassing fully-supervised baselines in a semi-supervised manner. Ablation study further proves the effectiveness of our method

. The source code is available at https://github.com/hangyu94/Ada-CM.
****

GanOrCon: Are Generative Models Useful for Few-Shot Segmentation?
Oindrila Saha, Zezhou Cheng, Subhransu Maji; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9991-10000
Advances in generative modeling based on GANs has motivated the community to fin d their use beyond image generation and editing tasks. In particular, several re cent works have shown that GAN representations can be re-purposed for discrimina tive tasks such as part segmentation, especially when training data is limited. But how do these improvements stack-up against recent advances in self-supervise d learning? Motivated by this we present an alternative approach based on contra stive learning and compare their performance on standard few-shot part segmentat ion benchmarks. Our experiments reveal that not only do the GAN-based approach o ffer no significant performance advantage, their multi-step training is complex, nearly an order-of-magnitude slower, and can introduce additional bias. These e xperiments suggest that the inductive biases of generative models, such as their ability to disentangle shape and texture, are well captured by standard feed-fo rward networks trained using contrastive learning.
****

Bi-Level Doubly Variational Learning for Energy-Based Latent Variable Models
Ge Kan, Jinhu Lü, Tian Wang, Baochang Zhang, Aichun Zhu, Lei Huang, Guodong Guo, Hichem Snoussi; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2022, pp. 18460-18469
Energy-based latent variable models (EBLVMs) are more expressive than convention al energy-based models. However, its potential on visual tasks are limited by it s training process based on maximum likelihood estimate that requires sampling f rom two intractable distributions. In this paper, we propose Bi-level doubly var iational learning (BiDVL), which is based on a new bi-level optimization framewo rk and two tractable variational distributions to facilitate learning EBLVMs. Pa rticularly, we lead a decoupled EBLVM consisting of a marginal energy-based dist ribution and a structural posterior to handle the difficulties when learning dee p EBLVMs on images. By choosing a symmetric KL divergence in the lower level of our framework, a compact BiDVL for visual tasks can be obtained. Our model achie ves impressive image generation performance over related works. It also demonstr ates the significant capacity of testing image reconstruction and out-of-distrib ution detection.
****

SplitNets: Designing Neural Architectures for Efficient Distributed Computing on Head-Mounted Systems
Xin Dong, Barbara De Salvo, Meng Li, Chiao Liu, Zhongnan Qu, H.T. Kung, Ziyun Li ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 12559-12569
We design deep neural networks (DNNs) and corresponding networks' splittings to distribute DNNs' workload to camera sensors and a centralized aggregator on head -mounted devices to meet system performance targets in inference accuracy and la tency under the given hardware resource constraints. To achieve an optimal balan ce among computation, communication, and performance, a split-aware neural archi tecture search framework, SplitNets, is introduced to conduct model designing, s plitting, and communication reduction simultaneously. We further extend the fram ework to multi-view systems for learning to fuse inputs from multiple camera sen sors with optimal performance and systemic efficiency. We validate SplitNets for single-view system on ImageNet as well as multi-view system on 3D classificatio n, and show that the SplitNets framework achieves state-of-the-art (SOTA) perfor mance and system latency compared with existing approaches.
****

Masked-Attention Mask Transformer for Universal Image Segmentation
Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, Rohit Girdha r; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2022, pp. 1290-1299
Image segmentation groups pixels with different semantics, e.g., category or ins

tance membership. Each choice of semantics defines a task. While only the semant
ics of each task differ, current research focuses on designing specialized archi
tectures for each task. We present Masked-attention Mask Transformer (Mask2Forme
r), a new architecture capable of addressing any image segmentation task (panopt
ic, instance or semantic). Its key components include masked attention, which ex
tracts localized features by constraining cross-attention within predicted mask
regions. In addition to reducing the research effort by at least three times, it
 outperforms the best specialized architectures by a significant margin on four
popular datasets. Most notably, Mask2Former sets a new state-of-the-art for pano
ptic segmentation (57.8 PQ on COCO), instance segmentation (50.1 AP on COCO) and
 semantic segmentation (57.7 mIoU on ADE20K).
****

Reading To Listen at the Cocktail Party: Multi-Modal Speech Separation
Akam Rahimi, Triantafyllos Afouras, Andrew Zisserman; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10493-
10502
The goal of this paper is speech separation and enhancement in multi-speaker and
 noisy environments using a combination of different modalities. Previous works
have shown good performance when conditioning on temporal or static visual evide
nce such as synchronised lip movements or face identity. In this paper we presen
t a unified framework for multi-modal speech separation and enhancement based on
 synchronous or asynchronous cues. To that end we make the following contributio
ns: (i) we design a modern Transformer-based architecture which inputs and outpu
ts raw waveforms and is tailored to fuse different modalities to solve the speec
h separation task; (ii) we propose conditioning on the text content of a sentenc
e alone or in combination with visual information; (iii) we demonstrate the robu
stness of our model to audio-visual synchronisation offsets; and, (iv) we obtain
 state-of-the art performance on the well-established benchmark datasets LRS2 an
d LRS3.
****

AxIoU: An Axiomatically Justified Measure for Video Moment Retrieval
Riku Togashi, Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Tetsuya Sak
ai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2022, pp. 21076-21085
Evaluation measures have a crucial impact on the direction of research. Therefor
e, it is of utmost importance to develop appropriate and reliable evaluation mea
sures for new applications where conventional measures are not well suited. Vide
o Moment Retrieval (VMR) is one such application, and the current practice is to
 use R@K,\theta for evaluating VMR systems. However, this measure has two disadv
antages. First, it is rank-insensitive: It ignores the rank positions of success
fully localised moments in the top-K ranked list by treating the list as a set.
Second, it binarises the Intersection over Union (IoU) of each retrieved video m
oment using the threshold \theta and thereby ignoring fine-grained localisation
quality of ranked moments. We propose an alternative measure for evaluating VMR,
 called Average Max IoU (AxIoU), which is free from the above two problems. We s
how that AxIoU satisfies two important axioms for VMR evaluation, namely, Invari
ance against Redundant Moments and Monotonicity with respect to the Best Moment,
 and also that R@K,\theta satisfies the first axiom only. We also empirically ex
amine how AxIoU agrees with R@K,\theta, as well as its stability with respect to
 change in the test data and human-annotated temporal boundaries.
****

NOC-REK: Novel Object Captioning With Retrieved Vocabulary From External Knowled
ge
Duc Minh Vo, Hong Chen, Akihiro Sugimoto, Hideki Nakayama; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1
8000-18008
Novel object captioning aims at describing objects absent from training data, wi
th the key ingredient being the provision of object vocabulary to the model. Alt
hough existing methods heavily rely on an object detection model, we view the de
tection step as vocabulary retrieval from an external knowledge in the form of e

mbeddings for any object's definition from Wiktionary, where we use in the retri eval image region features learned from a transformers model. We propose an end- to-end Novel Object Captioning with Retrieved vocabulary from External Knowledge method (NOC-REK), which simultaneously learns vocabulary retrieval and caption generation, successfully describing novel objects outside of the training datase t. Furthermore, our model eliminates the requirement for model retraining by sim ply updating the external knowledge whenever a novel object appears. Our compreh ensive experiments on held-out COCO and Nocaps datasets show that our NOC-REK is considerably effective against SOTAs.
****

Boosting Robustness of Image Matting With Context Assembling and Strong Data Aug mentation
Yutong Dai, Brian Price, He Zhang, Chunhua Shen; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11707-11716
Deep image matting methods have achieved increasingly better results on benchmar ks (e.g., Composition-1k/alphamatting.com). However, the robustness, including r obustness to trimaps and generalization to images from different domains, is sti ll under-explored. Although some works propose to either refine the trimaps or a dapt the algorithms to real-world images via extra data augmentation, none of th em has taken both into consideration, not to mention the significant performance deterioration on benchmarks while using those data augmentation. To fill this g ap, we propose an image matting method which achieves higher robustness (RMat) v ia multilevel context assembling and strong data augmentation targeting matting. Specifically, we first build a strong matting framework by modeling ample globa l information with transformer blocks in the encoder, and focusing on details in combination with convolution layers as well as a low-level feature assembling a ttention block in the decoder. Then, based on this strong baseline, we analyze c urrent data augmentation and explore simple but effective strong data augmentati on to boost the baseline model and contribute a more generalizable matting metho d. Compared with previous methods, the proposed method not only achieves state-o f-the-art results on the Composition-1k benchmark (11% improvement on SAD and 27 % improvement on Grad) with smaller model size, but also shows more robust gener alization results on other benchmarks, on real-world images, and also on varying coarse-to-fine trimaps with our extensive experiments.
****

Group R-CNN for Weakly Semi-Supervised Object Detection With Points
Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, Kai Chen; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 9417-9426
We study the problem of weakly semi-supervised object detection with points (WSS OD-P), where the training data is combined by a small set of fully annotated ima ges with bounding boxes and a large set of weakly-labeled images with only a sin gle point annotated for each instance. The core of this task is to train a point -to-box regressor on well-labeled images that can be used to predict credible bo unding boxes for each point annotation. We challenge the prior belief that exist ing CNN-based detectors are not compatible with this task. Based on the classic R-CNN architecture, we propose an effective point-to-box regressor: Group R-CNN. Group R-CNN first uses instance-level proposal grouping to generate a group of proposals for each point annotation and thus can obtain a high recall rate. To b etter distinguish different instances and improve precision, we propose instance -level proposal assignment to replace the vanilla assignment strategy adopted in original R-CNN methods. As naive instance-level assignment brings converging di fficulty, we propose instance-aware representation learning which consists of in stance-aware feature enhancement and instance-aware parameter generation to over come this issue. Comprehensive experiments on the MS-COCO benchmark demonstrate the effectiveness of our method. Specifically, Group R-CNN significantly outperf orms the prior method Point DETR by 3.9 mAP with 5% well-labeled images, which i s the most challenging scenario. The source code can be found at https://github. com/jshilong/GroupRCNN.
****

Weakly-Supervised Action Transition Learning for Stochastic Human Motion Prediction

Wei Mao, Miaomiao Liu, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8151-8160

We introduce the task of action-driven stochastic human motion prediction, which aims to predict multiple plausible future motions given a sequence of action labels and a short motion history. This differs from existing works, which predict motions that either do not respect any specific action category, or follow a single action label. In particular, addressing this task requires tackling two challenges: The transitions between the different actions must be smooth; the length of the predicted motion depends on the action sequence and varies significantly across samples. As we cannot realistically expect training data to cover sufficiently diverse action transitions and motion lengths, we propose an effective training strategy consisting of combining multiple motions from different actions and introducing a weak form of supervision to encourage smooth transitions. We then design a VAE-based model conditioned on both the observed motion and the action label sequence, allowing us to generate multiple plausible future motions of varying length. We illustrate the generality of our approach by exploring its use with two different temporal encoding models, namely RNNs and Transformers. Our approach outperforms baseline models constructed by adapting state-of-the-art single action-conditioned motion generation methods and stochastic human motion prediction approaches to our new task of action-driven stochastic motion prediction. Our code is available at https://github.com/wei-mao-2019/WAT.

****

Speech Driven Tongue Animation

Salvador Medina, Denis Tome, Carsten Stoll, Mark Tiede, Kevin Munhall, Alexander G. Hauptmann, Iain Matthews; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20406-20416

Advances in speech driven animation techniques allow the creation of convincing animations for virtual characters solely from audio data. Many existing approaches focus on facial and lip motion and they often do not provide realistic animation of the inner mouth. This paper addresses the problem of speech-driven inner mouth animation. Obtaining performance capture data of the tongue and jaw from video alone is difficult because the inner mouth is only partially observable during speech. In this work, we introduce a large-scale speech and mocap dataset that focuses on capturing tongue, jaw, and lip motion. This dataset enables research using data-driven techniques to generate realistic inner mouth animation from speech. We then propose a deep-learning based method for accurate and generalizable speech to tongue and jaw animation and evaluate several encoder-decoder network architectures and audio feature encoders. We find that recent self-supervised deep learning based audio feature encoders are robust, generalize well to unseen speakers and content, and work best for our task. To demonstrate the practical application of our approach, we show animations on high-quality parametric 3D face models driven by the landmarks generated from our speech-to-tongue animation method.

****

Hybrid Relation Guided Set Matching for Few-Shot Action Recognition

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19948-19957

Current few-shot action recognition methods reach impressive performance by learning discriminative features for each video via episodic training and designing various temporal alignment strategies. Nevertheless, they are limited in that (a) learning individual features without considering the entire task may lose the most relevant information in the current episode, and (b) these alignment strategies may fail in misaligned instances. To overcome the two limitations, we propose a novel Hybrid Relation guided Set Matching (HyRSM) approach that incorporates two key components: hybrid relation module and set matching metric. The purpose of the hybrid relation module is to learn task-specific embeddings by fully exploiting associated relations within and cross videos in an episode. Built upon

the task-specific features, we reformulate distance measure between query and su
pport videos as a set matching problem and further design a bidirectional Mean H
ausdorff Metric to improve the resilience to misaligned instances. By this means
, the proposed HyRSM can be highly informative and flexible to predict query cat
egories under the few-shot settings. We evaluate HyRSM on six challenging benchm
arks, and the experimental results show its superiority over the state-of-the-ar
t methods by a convincing margin. Project page: https://hyrsm-cvpr2022.github.io
/.
****

Self-Supervised Spatial Reasoning on Multi-View Line Drawings
Siyuan Xiang, Anbang Yang, Yanfei Xue, Yaoqing Yang, Chen Feng; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
pp. 12745-12754
Spatial reasoning on multi-view line drawings by state-of-the-art supervised dee
p networks is recently shown with puzzling low performances on the SPARE3D datas
et. Based on the fact that self-supervised learning is helpful when a large numb
er of data are available, we propose two self-supervised learning approaches to
improve the baseline performance for view consistency reasoning and camera pose
reasoning tasks on the SPARE3D dataset. For the first task, we use a self-superv
ised binary classification network to contrast the line drawing differences betw
een various views of any two similar 3D objects, enabling the trained networks t
o effectively learn detail-sensitive yet view-invariant line drawing representat
ions of 3D objects. For the second type of task, we propose a self-supervised mu
lti-class classification framework to train a model to select the correct corres
ponding view from which a line drawing is rendered. Our method is even helpful f
or the downstream tasks with unseen camera poses. Experiments show that our meth
ods could significantly increase the baseline performance in SPARE3D, while some
 popular self-supervised learning methods cannot.
****

Language-Bridged Spatial-Temporal Interaction for Referring Video Object Segment
ation
Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, Si Liu; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2022, pp. 4964-4973
Referring video object segmentation aims to predict foreground labels for object
s referred by natural language expressions in videos. Previous methods either de
pend on 3D ConvNets or incorporate additional 2D ConvNets as encoders to extract
 mixed spatial-temporal features. However, these methods suffer from spatial mis
alignment or false distractors due to delayed and implicit spatial-temporal inte
raction occurring in the decoding phase. To tackle these limitations, we propose
 a Language-Bridged Duplex Transfer (LBDT) module which utilizes language as an
intermediary bridge to accomplish explicit and adaptive spatial-temporal interac
tion earlier in the encoding phase. Concretely, cross-modal attention is perform
ed among the temporal encoder, referring words and the spatial encoder to aggreg
ate and transfer language-relevant motion and appearance information. In additio
n, we also propose a Bilateral Channel Activation (BCA) module in the decoding p
hase for further denoising and highlighting the spatial-temporal consistent feat
ures via channel-wise activation. Extensive experiments show our method achieves
 new state-of-the-art performances on four popular benchmarks with 6.8% and 6.9%
 absolute AP gains on A2D Sentences and J-HMDB Sentences respectively, while con
suming around 7x less computational overhead.
****

Cross-Patch Dense Contrastive Learning for Semi-Supervised Segmentation of Cellu
lar Nuclei in Histopathologic Images
Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, Jing Qin; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11666
-11675
We study the semi-supervised learning problem, using a few labeled data and a la
rge amount of unlabeled data to train the network, by developing a cross-patch d
ense contrastive learning framework, to segment cellular nuclei in histopatholog

ic images. This task is motivated by the expensive burden on collecting labeled data for histopathologic image segmentation tasks. The key idea of our method is to align features of teacher and student networks, sampled from cross-image in both patch- and pixel-levels, for enforcing the intra-class compactness and inter-class separability of features that as we shown is helpful for extracting valuable knowledge from unlabeled data. We also design a novel optimization framework that combines consistency regularization and entropy minimization techniques, showing good property in eviction of gradient vanishing. We assess the proposed method on two publicly available datasets, and obtain positive results on extensive experiments, outperforming the state-of-the-art methods. Codes are available at https://github.com/zzw-szu/CDCL.
****

Frame-Wise Action Representations for Long Videos via Sequence Contrastive Learning
Minghao Chen, Fangyun Wei, Chong Li, Deng Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13801-13810
Prior works on action representation learning mainly focus on designing various architectures to extract the global representations for short video clips. In contrast, many practical applications such as video alignment have strong demand for learning dense representations for long videos. In this paper, we introduce a novel contrastive action representation learning (CARL) framework to learn frame-wise action representations, especially for long videos, in a self-supervised manner. Concretely, we introduce a simple yet efficient video encoder that considers spatio-temporal context to extract frame-wise representations. Inspired by the recent progress of self-supervised learning, we present a novel sequence contrastive loss (SCL) applied on two correlated views obtained through a series of spatio-temporal data augmentations. SCL optimizes the embedding space by minimizing the KL-divergence between the sequence similarity of two augmented views and a prior Gaussian distribution of timestamp distance. Experiments on FineGym, PennAction and Pouring datasets show that our method outperforms previous state-of-the-art by a large margin for downstream fine-grained action classification. Surprisingly, although without training on paired videos, our approach also shows outstanding performance on video alignment and fine-grained frame retrieval tasks. Code and models will be made public.
****

Coarse-To-Fine Deep Video Coding With Hyperprior-Guided Mode Prediction
Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5921-5930
The previous deep video compression approaches only use the single scale motion compensation strategy and rarely adopt the mode prediction technique from the traditional standards like H.264/H.265 for both motion and residual compression. In this work, we first propose a coarse-to-fine (C2F) deep video compression framework for better motion compensation, in which we perform motion estimation, compression and compensation twice in a coarse to fine manner. Our C2F framework can achieve better motion compensation results without significantly increasing bit costs. Observing hyperprior information (i.e., the mean and variance values) from the hyperprior networks contains discriminant statistical information of different patches, we also propose two efficient hyperprior-guided mode prediction methods. Specifically, using hyperprior information as the input, we propose two mode prediction networks to respectively predict the optimal block resolutions for better motion coding and decide whether to skip residual information from each block for better residual coding without introducing additional bit cost while bringing negligible extra computation cost. Comprehensive experimental results demonstrate our proposed C2F video compression framework equipped with the new hyperprior-guided mode prediction methods achieves the state-of-the-art performance on HEVC, UVG and MCL-JCV datasets.
****

Generalized Binary Search Network for Highly-Efficient Multi-View Stereo
Zhenxing Mi, Chang Di, Dan Xu; Proceedings of the IEEE/CVF Conference on Compute

r Vision and Pattern Recognition (CVPR), 2022, pp. 12991-13000
Multi-view Stereo (MVS) with known camera parameters is essentially a 1D search problem within a valid depth range. Recent deep learning-based MVS methods typically densely sample depth hypotheses in the depth range, and then construct prohibitively memory-consuming 3D cost volumes for depth prediction. Although coarse-to-fine sampling strategies alleviate this overhead issue to a certain extent, the efficiency of MVS is still an open challenge. In this work, we propose a novel method for highly efficient MVS that remarkably decreases the memory footprint, meanwhile clearly advancing state-of-the-art depth prediction performance. We investigate what a search strategy can be reasonably optimal for MVS taking into account of both efficiency and effectiveness. We first formulate MVS as a binary search problem, and accordingly propose a generalized binary search network for MVS. Specifically, in each step, the depth range is split into 2 bins with extra 1 error tolerance bin on both sides. A classification is performed to identify which bin contains the true depth. We also design three mechanisms to respectively handle classification errors, deal with out-of-range samples and decrease the training memory. The new formulation makes our method only sample a very small number of depth hypotheses in each step, which is highly memory efficient, and also greatly facilitates quick training convergence. Experiments on competitive benchmarks show that our method achieves state-of-the-art accuracy with much less memory. Particularly, our method obtains an overall score of 0.289 on DTU dataset and tops the first place on challenging Tanks and Temples advanced dataset among all the learning-based methods. Our code will be released at https://github.com/MiZhenxing/GBi-Net.
****

SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation
Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21371-21382
Adapting to a continuously evolving environment is a safety-critical challenge inevitably faced by all autonomous-driving systems. Existing image- and video-based driving datasets, however, fall short of capturing the mutable nature of the real world. In this paper, we introduce the largest synthetic dataset for autonomous driving, SHIFT. It presents discrete and continuous shifts in cloudiness, rain and fog intensity, time of day, and vehicle and pedestrian density. Featuring a comprehensive sensor suite and annotations for several mainstream perception tasks, SHIFT allows to investigate how a perception systems' performance degrades at increasing levels of domain shift, fostering the development of continuous adaptation strategies to mitigate this problem and assessing the robustness and generality of a model. Our dataset and benchmark toolkit are publicly available at https://www.vis.xyz/shift.
****

Adaptive Hierarchical Representation Learning for Long-Tailed Object Detection
Banghuai Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2313-2322
General object detectors are always evaluated on hand-designed datasets, e.g., MS COCO and Pascal VOC, which tend to maintain balanced data distribution over different classes. However, it goes against the practical applications in the real world which suffer from a heavy class imbalance problem, known as the long-tailed object detection. In this paper, we propose a novel method, named Adaptive Hierarchical Representation Learning (AHRL), from a metric learning perspective to address long-tailed object detection. We visualize each learned class representation in the feature space, and observe that some classes, especially under-represented scarce classes, are prone to cluster with analogous ones due to the lack of discriminative representation. Inspired by this, we propose to split the whole feature space into a hierarchical structure and eliminate the problem in a divide-and-conquer way. AHRL contains a two-stage training paradigm. First, we train a normal baseline model and construct the hierarchical structure under the unsupervised clustering method. Then, we design an AHR loss that consists of two optimization objectives. On the one hand, AHR loss retains the hierarchical struc

ture and keeps representation clusters away from each other. On the other hand, AHR loss adopts adaptive margins according to specific class pairs in the same cluster to further optimize locally. We conduct extensive experiments on the challenging LVIS dataset and AHRL outperforms all the existing state-of-the-art(SOTA) methods, with 29.1% segmentation AP and 29.3% box AP on LVIS v0.5 and 27.6% segmentation AP and 28.7% box AP on LVIS v1.0 based on ResNet-101. We hope our simple yet effective approach will serve as a solid baseline to help stimulate future research in long-tailed object detection. Code will be released soon.
****

FlexIT: Towards Flexible Semantic Image Translation
Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18270-18279
Deep generative models, like GANs, have considerably improved the state of the art in image synthesis, and are able to generate near photo-realistic images in structured domains such as human faces. Based on this success, recent work on image editing proceeds by projecting images to the GAN latent space and manipulating the latent vector. However, these approaches are limited in that only images from a narrow domain can be transformed, and with only a limited number of editing operations. We propose FlexIT, a novel method which can take any input image and a user-defined text instruction for editing. Our method achieves flexible and natural editing, pushing the limits of semantic image translation. First, FlexIT combines the input image and text into a single target point in the CLIP multimodal embedding space. Via the latent space of an autoencoder, we iteratively transform the input image toward the target point, ensuring coherence and quality with a variety of novel regularization terms. We propose an evaluation protocol for semantic image translation, and thoroughly evaluate our method on ImageNet. Code will be available at https://github.com/facebookresearch/SemanticImageTranslation/.
****

Face2Exp: Combating Data Biases for Facial Expression Recognition
Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, Bo Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20291-20300
Facial expression recognition (FER) is challenging due to the class imbalance caused by data collection. Existing studies tackle the data bias problem using only labeled facial expression dataset. Orthogonal to existing FER methods, we propose to utilize large unlabeled face recognition (FR) datasets to enhance FER. However, this raises another data bias problem---the distribution mismatch between FR and FER data. To combat the mismatch, we propose the Meta-Face2Exp framework, which consists of a base network and an adaptation network. The base network learns prior expression knowledge on class-balanced FER data while the adaptation network is trained to fit the pseudo labels of FR data generated by the base model. To combat the mismatch between FR and FER data, Meta-Face2Exp uses a circuit feedback mechanism, which improves the base network with the feedback from the adaptation network. Experiments show that our Meta-Face2Exp achieves comparable accuracy to state-of-the-art FER methods with 10% of the labeled FER data utilized by the baselines. We also demonstrate that the circuit feedback mechanism successfully eliminates data bias.
****

SAR-Net: Shape Alignment and Recovery Network for Category-Level 6D Object Pose and Size Estimation
Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, Xiangyang Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6707-6717
Given a single scene image, this paper proposes a method of Category-level 6D Object Pose and Size Estimation (COPSE) from the point cloud of the target object, without external real pose-annotated training data. Specifically, beyond the visual cues in RGB images, we rely on the shape information predominately from the depth (D) channel. The key idea is to explore the shape alignment of each insta

nce against its corresponding category-level template shape, and the symmetric c orrespondence of each object category for estimating a coarse 3D object shape. O ur framework deforms the point cloud of the category-level template shape to ali gn the observed instance point cloud for implicitly representing its 3D rotation . Then we model the symmetric correspondence by predicting symmetric point cloud from the partially observed point cloud. The concatenation of the observed poin t cloud and symmetric one reconstructs a coarse object shape, thus facilitating object center (3D translation) and 3D size estimation. Extensive experiments on the category-level NOCS benchmark demonstrate that our lightweight model still c ompetes with state-of-the-art approaches that require labeled real-world images. We also deploy our approach to a physical Baxter robot to perform grasping task s on unseen but category-known instances, and the results further validate the e fficacy of our proposed model. Code and pre-trained models are available on the project webpage.

****

Whose Hands Are These? Hand Detection and Hand-Body Association in the Wild

Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 4889-4899

We study a new problem of detecting hands and finding the location of the corres ponding person for each detected hand. This task is helpful for many downstream tasks such as hand tracking and hand contact estimation. Associating hands with people is challenging in unconstrained conditions since multiple people can be p resent in the scene with varying overlaps and occlusions. We propose a novel end -to-end trainable convolutional network that can jointly detect hands and the bo dy location for the corresponding person. Our method first detects a set of hand s and bodies and uses a novel Hand-Body Association Network to predict associati on scores between them. We use these association scores to find the body locatio n for each detected hand. We also introduce a new challenging dataset called Bod yHands containing unconstrained images with hand and their corresponding body lo cations annotations. We conduct extensive experiments on BodyHands and another p ublic dataset to show the effectiveness of our method. Finally, we demonstrate t he benefits of hand-body association in two critical applications: hand tracking and hand contact estimation. Our experiments show that hand tracking and hand c ontact estimation methods can be improved significantly by reasoning about the h and-body association.

****

Mega-NERF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs

Haithem Turki, Deva Ramanan, Mahadev Satyanarayanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12922-1 2931

We use neural radiance fields (NeRFs) to build interactive 3D environments from large-scale visual captures spanning buildings or even multiple city blocks coll ected primarily from drones. In contrast to single object scenes (on which NeRFs are traditionally evaluated), our scale poses multiple challenges including (1) the need to model thousands of images with varying lighting conditions, each of which capture only a small subset of the scene, (2) prohibitively large model c apacities that make it infeasible to train on a single GPU, and (3) significant challenges for fast rendering that would enable interactive fly-throughs. To add ress these challenges, we begin by analyzing visibility statistics for large-sca le scenes, motivating a sparse network structure where parameters are specialize d to different regions of the scene. We introduce a simple geometric clustering algorithm for data parallelism that partitions training images (or rather pixels) into different NeRF submodules that can be trained in parallel. We evaluate ou r approach on existing datasets (Quad 6k and UrbanScene3D) as well as against ou r own drone footage, improving training speed by 3x and PSNR by 12%. We also eva luate recent NeRF fast renderers on top of Mega-NeRF and introduce a novel metho d that exploits temporal coherence. Our technique achieves a 40x speedup over co nventional NeRF rendering while remaining within 0.8 db in PSNR quality, exceedi ng the fidelity of existing fast renderers.

****
PINA: Learning a Personalized Implicit Neural Avatar From a Single RGB-D Video S equence
Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, Otmar Hilliges; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2022, pp. 20470-20480
We present a novel method to learn Personalized Implicit Neural Avatars (PINA) f rom a short RGB-D sequence. This allows non-expert users to create a detailed an d personalized virtual copy of themselves, which can be animated with realistic clothing deformations. PINA does not require complete scans, nor does it require a prior learned from large datasets of clothed humans. Learning a complete avat ar in this setting is challenging, since only few depth observations are availab le, which are noisy and incomplete (i.e. only partial visibility of the body per frame). We propose a method to learn the shape and non-rigid deformations via a pose-conditioned implicit surface and a deformation field, defined in canonical space. This allows us to fuse all partial observations into a single consistent canonical representation. Fusion is formulated as a global optimization problem over the pose, shape and skinning parameters. The method can learn neural avata rs from real noisy RGB-D sequences for a diverse set of people and clothing styl es and these avatars can be animated given unseen motion sequences.
****
Forecasting From LiDAR via Future Object Detection
Neehar Peri, Jonathon Luiten, Mengtian Li, Aljoša Ošep, Laura Leal-Taixé, Deva R amanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 17202-17211
Object detection and forecasting are fundamental components of embodied percepti on. These two problems, however, are largely studied in isolation by the communi ty. In this paper, we propose an end-to-end approach for motion forecasting base d on raw sensor measurement as opposed to ground truth tracks. Instead of predic ting the current frame locations and forecast forward in time, we directly predi ct future object locations and backcast in time to determine where each trajecto ry began. Our approach not only greatly boosts the overall accuracy compared to modular or other end-to-end baselines, it also prompts us to rethink the role of explicit tracking for embodied perception. Additionally, by linking future and current locations in a multiple-to-one manner, our approach is able to reason ab out multiple futures, a capability that was previously considered difficult for end-to-end approaches. We conduct extensive experiments on the popular autonomou s driving dataset nuScenes and demonstrate the empirical effectiveness of our ap proach. In addition, we investigate the appropriateness of reusing standard fore casting metrics for an end-to-end setup, and find a number of flaws which allow us to build simple baselines to game these metrics. We address this issue with a novel set of joint forecasting and detection metrics that extend the commonly u sed AP metrics used in the detection community to measuring forecasting accuracy .
****
CRAFT: Cross-Attentional Flow Transformer for Robust Optical Flow
Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, Hongy uan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 17602-17611
Optical flow estimation aims to find the 2D motion field by identifying correspo nding pixels between two images. Despite the tremendous progress of deep learnin g-based optical flow methods, it remains a challenge to accurately estimate larg e displacements with motion blur. This is mainly because the correlation volume, the basis of pixel matching, is computed as the dot product of the convolutiona l features of the two images. The locality of convolutional features makes the c omputed correlations susceptible to various noises. On large displacements with motion blur, noisy correlations could cause severe errors in the estimated flow. To overcome this challenge, we propose a new architecture "CRoss-Attentional Fl ow Transformer" (CRAFT), aiming to revitalize the correlation volume computation . In CRAFT, a Semantic Smoothing Transformer layer transforms the features of on

e frame, making them more global and semantically stable. In addition, the dot-p roduct correlations are replaced with transformer Cross-Frame Attention. This la yer filters out feature noises through the Query and Key projections, and comput es more accurate correlations. On Sintel (Final) and KITTI (foreground) benchmar ks, CRAFT has achieved new state-of-the-art performance. Moreover, to test the r obustness of different models on large motions, we designed an image shifting at tack that shifts input images to generate large artificial motions. Under this a ttack, CRAFT performs much more robustly than two representative methods, RAFT a nd GMA. The code of CRAFT is is available at https://github.com/askerlee/craft.
****

Adversarial Eigen Attack on Black-Box Models
Linjun Zhou, Peng Cui, Xingxuan Zhang, Yinan Jiang, Shiqiang Yang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 15254-15262
Black-box adversarial attack has aroused much research attention for its difficu lty on nearly no available information of the attacked model and the additional constraint on the query budget. A common way to improve attack efficiency is to transfer the gradient information of a white-box substitute model trained on an extra dataset. In this paper, we deal with a more practical setting where a pre-trained white-box model with network parameters is provided without extra traini ng data. To solve the model mismatch problem between the white-box and black-box models, we propose a novel algorithm EigenBA by systematically integrating grad ient-based white-box method and zeroth-order optimization in black-box methods. We theoretically show the optimal directions of perturbations for each step are closely related to the right singular vectors of the Jacobian matrix of the pret rained white-box model. Extensive experiments on ImageNet, CIFAR-10 and WebVisio n show that EigenBA can consistently and significantly outperform state-of-the-a rt baselines in terms of success rate and attack efficiency.
****

Training Quantised Neural Networks With STE Variants: The Additive Noise Anneali ng Algorithm
Matteo Spallanzani, Gian Paolo Leonardi, Luca Benini; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 470-47 9
Training quantised neural networks (QNNs) is a non-differentiable optimisation p roblem since weights and features are output by piecewise constant functions. Th e standard solution is to apply the straight-through estimator (STE), using diff erent functions during the inference and gradient computation steps. Several STE variants have been proposed in the literature aiming to maximise the task accur acy of the trained network. In this paper, we analyse STE variants and study the ir impact on QNN training. We first observe that most such variants can be model led as stochastic regularisations of stair functions; although this intuitive in terpretation is not new, our rigorous discussion generalises to further variants . Then, we analyse QNNs mixing different regularisations, finding that some suit ably synchronised smoothing of each layer map is required to guarantee pointwise compositional convergence to the target discontinuous function. Based on these theoretical insights, we propose additive noise annealing (ANA), a new algorithm to train QNNs encompassing standard STE and its variants as special cases. When testing ANA on the CIFAR-10 image classification benchmark, we find that the ma jor impact on task accuracy is not due to the qualitative shape of the regularis ations but to the proper synchronisation of the different STE variants used in a network, in accordance with the theoretical results.
****

Split Hierarchical Variational Compression
Tom Ryder, Chen Zhang, Ning Kang, Shifeng Zhang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 386-395
Variational autoencoders (VAEs) have witnessed great success in performing the c ompression of image datasets. This success, made possible by the bits-back codin g framework, has produced competitive compression performance across many benchm arks. However, despite this, VAE architectures are currently limited by a combin

ation of coding practicalities and compression ratios. That is, not only do stat e-of-the-art methods, such as normalizing flows, often demonstrate out-performan ce, but the initial bits required in coding makes single and parallel image comp ression challenging. To remedy this, we introduce Split Hierarchical Variational Compression (SHVC). SHVC introduces two novelties. Firstly, we propose an effic ient autoregressive prior, the autoregressive sub-pixel convolution, that allows a generalisation between per-pixel autoregressions and fully factorised probabi lity models. Secondly, we define our coding framework, the autoregressive initia l bits, that flexibly supports parallel coding and avoids -- for the first time -- many of the practicalities commonly associated with bits-back coding. In our experiments, we demonstrate SHVC is able to achieve state-of-the-art compression performance across full-resolution lossless image compression tasks, with up to 100x fewer model parameters than competing VAE approaches.
****

Video Swin Transformer
Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, Han Hu; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3202-3211
The vision community is witnessing a modeling shift from CNNs to Transformers, w here pure Transformer architectures have attained top accuracy on the major vide o recognition benchmarks. These video models are all built on Transformer layers that globally connect patches across the spatial and temporal dimensions. In th is paper, we instead advocate an inductive bias of locality in video Transformer s, which leads to a better speed-accuracy trade-off compared to previous approac hes which compute self-attention globally even with spatial-temporal factorizati on. The locality of the proposed video architecture is realized by adapting the Swin Transformer designed for the image domain, while continuing to leverage the power of pre-trained image models. Our approach achieves state-of-the-art accur acy on a broad range of video recognition benchmarks, including on action recogn ition (84.9 top-1 accuracy on Kinetics-400 and 85.9 top-1 accuracy on Kinetics-6 00 with ~20xless pre-training data and ~3xsmaller model size) and temporal model ing (69.6 top-1 accuracy on Something-Something v2).
****

Privacy Preserving Partial Localization
Marcel Geppert, Viktor Larsson, Johannes L. Schönberger, Marc Pollefeys; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17337-17347
Recently proposed privacy preserving solutions for cloud-based localization rely on lifting traditional point-based maps to randomized 3D line clouds. While the lifted representation is effective in concealing private information, there are two fundamental limitations. First, without careful construction of the line cl ouds, the representation is vulnerable to density-based inversion attacks. Secon dly, after successful localization, the precise camera orientation and position is revealed to the server. However, in many scenarios, the pose itself might be sensitive information. We propose a principled approach overcoming these limitat ions, based on two observations. First, a full 6 DoF pose is not always necessar y, and in combination with egomotion tracking even a one dimensional localizatio n can reduce uncertainty and correct drift. Secondly, by lifting to parallel pla nes instead of lines, the map only provides partial constraints on the query pos e, preventing the server from knowing the exact query location. If the client re quires a full 6 DoF pose, it can be obtained by fusing the result from multiple queries, which can be temporally and spatially disjoint. We demonstrate the prac tical feasibility of this approach and show a small performance drop compared to both the conventional and privacy preserving approaches.
****

Cross-Modal Background Suppression for Audio-Visual Event Localization
Yan Xia, Zhou Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 19989-19998
Audiovisual Event (AVE) localization requires the model to jointly localize an e vent by observing audio and visual information. However, in unconstrained videos

, both information types may be inconsistent or suffer from severe background no
ise. Hence this paper proposes a novel cross-modal background suppression networ
k for AVE task, operating at the time- and event-level, aiming to improve locali
zation performance through suppressing asynchronous audiovisual background frame
s from the examined events and reducing redundant noise. Specifically, the time-
level background suppression scheme forces the audio and visual modality to focu
s on the related information in the temporal dimension that the opposite modalit
y considers essential, and reduces attention to the segments that the other moda
l considers as background. The event-level background suppression scheme uses th
e class activation sequences predicted by audio and visual modalities to control
 the final event category prediction, which can effectively suppress noise event
s occurring accidentally in a single modality. Furthermore, we introduce a cross
-modal gated attention scheme to extract relevant visual regions from complex sc
enes exploiting both global visual and audio signals. Extensive experiments show
 our method outperforms the state-of-the-art methods by a large margin in both s
upervised and weakly supervised AVE settings.
****

Mutual Quantization for Cross-Modal Search With Noisy Labels
Erkun Yang, Dongren Yao, Tongliang Liu, Cheng Deng; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7551-756
0
Deep cross-modal hashing has become an essential tool for supervised multimodal
search. These models tend to be optimized with large, curated multimodal dataset
s, where most labels have been manually verified. Unfortunately, in many scenari
os, such accurate labeling may not be available. In contrast, datasets with low-
quality annotations may be acquired, which inevitably introduce numerous mistake
s or label noise and therefore degrade the search performance. To address the ch
allenge, we present a general robust cross-modal hashing framework to correlate
distinct modalities and combat noisy labels simultaneously. More specifically, w
e propose a proxy-based contrastive (PC) loss to mitigate the gap between differ
ent modalities and train networks for different modalities jointly with small-lo
ss samples that are selected with the PC loss and a mutual quantization loss. Th
e small-loss sample selection from such joint loss can help choose confident exa
mples to guide the model training, and the mutual quantization loss can maximize
 the agreement between different modalities and is beneficial to improve the eff
ectiveness of sample selection. Experiments on three widely-used multimodal data
sets show that our method significantly outperforms existing state-of-the-art me
thods.
****

Lagrange Motion Analysis and View Embeddings for Improved Gait Recognition
Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, Yunhong Wang; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 20249-20258
Gait is considered the walking pattern of human body, which includes both shape
and motion cues. However, the main-stream appearance-based methods for gait reco
gnition rely on the shape of silhouette. It is unclear whether motion can be exp
licitly represented in the gait sequence modeling. In this paper, we analyzed hu
man walking using the Lagrange's equation and come to the conclusion that second
-order information in the temporal dimension is necessary for identification. We
 designed a second-order motion extraction module based on the conclusions drawn
. Also, a light weight view-embedding module is designed by analyzing the proble
m that current methods to cross-view task do not take view itself into considera
tion explicitly. Experiments on CASIA-B and OU-MVLP datasets show the effectiven
ess of our method and some visualization for extracted motion are done to show t
he interpretability of our motion extraction module.
****

SphereSR: 360deg Image Super-Resolution With Arbitrary Projection via Continuous
 Spherical Image Representation
Youngho Yoon, Inchul Chung, Lin Wang, Kuk-Jin Yoon; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5677-568

6
The 360deg imaging has recently gained much attention; however, its angular resolution is relatively lower than that of a narrow field-of-view (FOV) perspective image as it is captured using a fisheye lens with the same sensor size. Therefore, it is beneficial to super-resolve a 360deg image. Several attempts have been made, but mostly considered equirectangular projection (ERP) as one of the ways for 360deg image representation despite the latitude-dependent distortions. In that case, as the output high-resolution (HR) image is always in the same ERP format as the lowresolution (LR) input, additional information loss may occur when transforming the HR image to other projection types. In this paper, we propose SphereSR, a novel framework to generate a continuous spherical image representation from an LR 360deg image, with the goal of predicting the RGB values at given spherical coordinates for superresolution with an arbitrary 360deg image projection. Specifically, first we propose a feature extraction module that represents the spherical data based on an icosahedron and that efficiently extracts features on the spherical surface. We then propose a spherical local implicit image function (SLIIF) to predict RGB values at the spherical coordinates. As such, SphereSR flexibly reconstructs an HR image given an arbitrary projection type. Experiments on various benchmark datasets show that the proposed method significantly surpasses existing methods in terms of performance.
****

Neural Mesh Simplification
Rolandos Alexandros Potamias, Stylianos Ploumpis, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18583-18592
Despite the advent in rendering, editing and preprocessing methods of 3D meshes, their real-time execution remains still infeasible for large-scale meshes. To ease and accelerate such processes, mesh simplification methods have been introduced with the aim to reduce the mesh resolution while preserving its appearance. In this work we attempt to tackle the novel task of learnable and differentiable mesh simplification. Compared to traditional simplification approaches that collapse edges in a greedy iterative manner, we propose a fast and scalable method that simplifies a given mesh in one-pass. The proposed method unfolds in three steps. Initially, a subset of the input vertices is sampled using a sophisticated extension of random sampling. Then, we train a sparse attention network to propose candidate triangles based on the edge connectivity of the sampled vertices. Finally, a classification network estimates the probability that a candidate triangle will be included in the final mesh. The fast, lightweight and differentiable properties of the proposed method makes it possible to be plugged in every learnable pipeline without introducing a significant overhead. We evaluate both the sampled vertices and the generated triangles under several appearance error measures and compare its performance against several state-of-the-art baselines. Furthermore, we showcase that the running performance can be up to 10-times faster than traditional methods.
****

Cloth-Changing Person Re-Identification From a Single Image With Gait Prediction and Regularization
Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14278-14287
Cloth-Changing person re-identification (CC-ReID) aims at matching the same person across different locations over a long-duration, e.g., over days, and therefore inevitably has cases of changing clothing. In this paper, we focus on handling well the CC-ReID problem under a more challenging setting, i.e., just from a single image, which enables an efficient and latency-free person identity matching for surveillance. Specifically, we introduce Gait recognition as an auxiliary task to drive the Image ReID model to learn cloth-agnostic representations by leveraging personal unique and cloth-independent gait information, we name this framework as GI-ReID. GI-ReID adopts a two-stream architecture that consists of an image ReID-Stream and an auxiliary gait recognition stream (Gait-Stream). The G

ait-Stream, that is discarded in the inference for high efficiency, acts as a re gulator to encourage the ReID-Stream to capture cloth-invariant biometric motion features during the training. To get temporal continuous motion cues from a sin gle image, we design a Gait Sequence Prediction (GSP) module for Gait-Stream to enrich gait information. Finally, a semantics consistency constraint over two st reams is enforced for effective knowledge regularization. Extensive experiments on multiple image-based Cloth-Changing ReID benchmarks, e.g., LTCC, PRCC, Real28 , and VC-Clothes, demonstrate that GI-ReID performs favorably against the state- of-the-art methods.
****

BoxeR: Box-Attention for 2D and 3D Transformers
Duy-Kien Nguyen, Jihong Ju, Olaf Booij, Martin R. Oswald, Cees G. M. Snoek; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 4773-4782
In this paper, we propose a simple attention mechanism, we call Box-Attention. I t enables spatial interaction between grid features, as sampled from boxes of in terest, and improves the learning capability of transformers for several vision tasks. Specifically, we present BoxeR, short for Box Transformer, which attends to a set of boxes by predicting their transformation from a reference window on an input feature map. The BoxeR computes attention weights on these boxes by con sidering its grid structure. Notably, BoxeR-2D naturally reasons about box infor mation within its attention module, making it suitable for end-to-end instance d etection and segmentation tasks. By learning invariance to rotation in the box-a ttention module, BoxeR-3D is capable of generating discriminative information fr om a bird's-eye view plane for 3D end-to-end object detection. Our experiments d emonstrate that the proposed BoxeR-2D achieves state-of-the-art results on COCO detection and instance segmentation. Besides, BoxeR-3D improves over the end-to- end 3D object detection baseline and already obtains a compelling performance fo r the vehicle category of Waymo Open, without any class-specific optimization. C ode is available at https://github.com/kienduynguyen/BoxeR.
****

Neural Architecture Search With Representation Mutual Information
Xiawu Zheng, Xiang Fei, Lei Zhang, Chenglin Wu, Fei Chao, Jianzhuang Liu, Wei Ze ng, Yonghong Tian, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2022, pp. 11912-11921
Performance evaluation strategy is one of the most important factors that determ ine the effectiveness and efficiency in Neural Architecture Search (NAS). Existi ng strategies, such as employing standard training or performance predictor, oft en suffer from high computational complexity and low generality. To address this issue, we propose to rank architectures by Representation Mutual Information (R MI). Specifically, given an arbitrary architecture that has decent accuracy, arc hitectures that have high RMI with it always yield good accuracies. As an accura te performance indicator to facilitate NAS, RMI not only generalizes well to dif ferent search spaces, but is also efficient enough to evaluate architectures usi ng only one batch of data. Building upon RMI, we further propose a new search al gorithm termed RMI-NAS, facilitating with a theorem to guarantee the global opti mal of the searched architecture. In particular, RMI-NAS first randomly samples architectures from the search space, which are then effectively classified as po sitive or negative samples by RMI. We then use these samples to train a random f orest to explore new regions, while keeping track of the distribution of positiv e architectures. When the sample size is sufficient, the architecture with the l argest probability from the aforementioned distribution is selected, which is th eoretically proved to be the optimal solution. The architectures searched by our method achieve remarkable top-1 accuracies with the magnitude times faster sear ch process. Besides, RMI-NAS also generalizes to different datasets and search s paces. Our code has been made available at https://git.openi.org.cn/PCL_AutoML/X NAS.
****

Deep Hyperspectral-Depth Reconstruction Using Single Color-Dot Projection
Chunyu Li, Yusuke Monno, Masatoshi Okutomi; Proceedings of the IEEE/CVF Conferen

ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19770-19779
Depth reconstruction and hyperspectral reflectance reconstruction are two active research topics in computer vision and image processing. Conventionally, these two topics have been studied separately using independent imaging setups and there is no existing method which can acquire depth and spectral reflectance simultaneously in one shot without using special hardware. In this paper, we propose a novel single-shot hyperspectral-depth reconstruction method using an off-the-shelf RGB camera and projector. Our method is based on a single color-dot projection, which simultaneously acts as structured light for depth reconstruction and spatially-varying color illuminations for hyperspectral reflectance reconstruction. To jointly reconstruct the depth and the hyperspectral reflectance from a single color-dot image, we propose a novel end-to-end network architecture that effectively incorporates a geometric color-dot pattern loss and a photometric hyperspectral reflectance loss. Through the experiments, we demonstrate that our hyperspectral-depth reconstruction method outperforms the combination of an existing state-of-the-art single-shot hyperspectral reflectance reconstruction method and depth reconstruction method.
****

M3T: Three-Dimensional Medical Image Classifier Using Multi-Plane and Multi-Slice Transformer
Jinseong Jang, Dosik Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20718-20729
In this study, we propose a three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer (M3T) network to classify Alzheimer's disease (AD) in 3D MRI images. The proposed network synergically combines 3D CNN, 2D CNN, and Transformer for accurate AD classification. The 3D CNN is used to perform natively 3D representation learning, while 2D CNN is used to utilize the pre-trained weights on large 2D databases and 2D representation learning. It is possible to efficiently extract the locality information for AD-related abnormalities in the local brain using CNN networks with inductive bias. The transformer network is also used to obtain attention relationships among multi-plane (axial, coronal, and sagittal) and multi-slice images after CNN. It is also possible to learn the abnormalities distributed over the wider region in the brain using the transformer without inductive bias. In this experiment, we used a training dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) which contains a total of 4,786 3D T1-weighted MRI images. For the validation data, we used dataset from three different institutions: The Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), The Open Access Series of Imaging Studies (OASIS), and some set of ADNI data independent from the training dataset. Our proposed M3T is compared to conventional 3D classification networks based on an area under the curve (AUC) and classification accuracy for AD classification. This study represents that the proposed network M3T achieved the highest performance in multi-institutional validation database, and demonstrates the feasibility of the method to efficiently combine CNN and Transformer for 3D medical images.
****

3MASSIV: Multilingual, Multimodal and Multi-Aspect Dataset of Social Media Short Videos
Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, Dinesh Manocha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21064-21075
We present 3MASSIV, a multilingual, multimodal and multi-aspect, expertly-annotated dataset of diverse short videos extracted from a social media platform. 3MASSIV comprises of 50k short videos (20 seconds average duration) and 100K unlabeled videos in 11 different languages and captures popular short video trends like pranks, fails, romance, comedy expressed via unique audio-visual formats like self-shot videos, reaction videos, lip-synching, self-sung songs, etc. 3MASSIV presents an opportunity for multimodal and multilingual semantic understanding on these unique videos by annotating them for concepts, affective states, media types, and audio language. We present a thorough analysis of 3MASSIV and highlight the variety and unique aspects of our dataset compared to other contemporary pop

ular datasets with strong baselines. We also show how the social media content i
n 3MASSIV is dynamic and temporal in nature which can be used for various semant
ic understanding tasks and cross-lingual analysis.
****

Can Neural Nets Learn the Same Model Twice? Investigating Reproducibility and Do
uble Descent From the Decision Boundary Perspective
Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richar
d Baraniuk, Micah Goldblum, Tom Goldstein; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13699-13708
We discuss methods for visualizing neural network decision boundaries and decisi
on regions. We use these visualizations to investigate issues related to reprodu
cibility and generalization in neural network training. We observe that changes
in model architecture (and its associate inductive bias) cause visible changes i
n decision boundaries, while multiple runs with the same architecture yield resu
lts with strong similarities, especially in the case of wide architectures. We a
lso use decision boundary methods to visualize double descent phenomena. We see
that decision boundary reproducibility depends strongly on model width. Near the
 threshold of interpolation, neural network decision boundaries become fragmente
d into many small decision regions, and these regions are non-reproducible. Mean
while, very narrows and very wide networks have high levels of reproducibility i
n their decision boundaries with relatively few decision regions. We discuss how
 our observations relate to the theory of double descent phenomena in convex mod
els. Code is available at https://github.com/somepago/dbViz.
****

Cross Domain Object Detection by Target-Perceived Dual Branch Distillation
Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei W
u, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2022, pp. 9570-9580
Cross domain object detection is a realistic and challenging task in the wild. I
t suffers from performance degradation due to large shift of data distributions
and lack of instance-level annotations in the target domain. Existing approaches
 mainly focus on either of these two difficulties, even though they are closely
coupled in cross domain object detection. To solve this problem, we propose a no
vel Target-perceived Dual-branch Distillation (TDD) framework. By integrating de
tection branches of both source and target domains in a unified teacher-student
learning scheme, it can reduce domain shift and generate reliable supervision ef
fectively. In particular, we first introduce a distinct Target Proposal Perceive
r between two domains. It can adaptively enhance source detector to perceive obj
ects in a target image, by leveraging target proposal contexts from iterative cr
oss-attention. Afterwards, we design a concise Dual Branch Self Distillation str
ategy for model training, which can progressively integrate complementary object
 knowledge from different domains via self-distillation in two branches. Finally
, we conduct extensive experiments on a number of widely-used scenarios in cross
 domain object detection. The results show that our TDD significantly outperform
s the state-of-the-art methods on all the benchmarks. The codes and models will
be released afterwards.
****

A Proposal-Based Paradigm for Self-Supervised Sound Source Localization in Video
s
Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, Xavier Alameda-Pineda; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 1029-1038
Humans can easily recognize where and how the sound is produced via watching a s
cene and listening to corresponding audio cues. To achieve such cross-modal perc
eption on machines, existing methods only use the maps generated by interpolatio
n operations to localize the sound source. As semantic object-level localization
 is more attractive for potential practical applications, we argue that these ex
isting map-based approaches only provide a coarse-grained and indirect descripti
on of the sound source. In this paper, we advocate a novel proposal-based paradi
gm that can directly perform semantic object-level localization, without any man

ual annotations. We incorporate the global response map as an unsupervised spati al constraint to weight the proposals according to how well they cover the estim ated global shape of the sound source. As a result, our proposal-based sound sou rce localization can be cast into a simpler Multiple Instance Learning (MIL) pro blem by filtering those instances corresponding to large sound-unrelated regions . Our method achieves state-of-the-art (SOTA) performance when compared to sever al baselines on multiple datasets.
****

Overcoming Catastrophic Forgetting in Incremental Object Detection via Elastic R esponse Distillation
Tao Feng, Mang Wang, Hangjie Yuan; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2022, pp. 9427-9436
Traditional object detectors are ill-equipped for incremental learning. However, fine-tuning directly on a well-trained detection model with only new data will lead to catastrophic forgetting. Knowledge distillation is a flexible way to mit igate catastrophic forgetting. In Incremental Object Detection (IOD), previous w ork mainly focuses on distilling for the combination of features and responses. However, they under-explore the information that contains in responses. In this paper, we propose a response-based incremental distillation method, dubbed Elast ic Response Distillation (ERD), which focuses on elastically learning responses from the classification head and the regression head. Firstly, our method transf ers category knowledge while equipping student detector with the ability to reta in localization information during incremental learning. In addition, we further evaluate the quality of all locations and provide valuable responses by the Ela stic Response Selection (ERS) strategy. Finally, we elucidate that the knowledge from different responses should be assigned with different importance during in cremental distillation. Extensive experiments conducted on MS COCO demonstrate t he proposed method achieves state-of-the-art performance, which substantially na rrows the performance gap towards full training.
****

GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction With R elational Reasoning
Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, Siheng Chen; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6 498-6507
Demystifying the interactions among multiple agents from their past trajectories is fundamental to precise and interpretable trajectory prediction. However, pre vious works only consider pair-wise interactions with limited relational reasoni ng. To promote more comprehensive interaction modeling for relational reasoning, we propose GroupNet, a multiscale hypergraph neural network, which is novel in terms of both interaction capturing and representation learning. From the aspect of interaction capturing, we propose a trainable multiscale hypergraph to captu re both pair-wise and group-wise interactions at multiple group sizes. From the aspect of interaction representation learning, we propose a three-element format that can be learnt end-to-end and explicitly reason some relational factors inc luding the interaction strength and category. We apply GroupNet into both CVAE-b ased prediction system and previous state-of-the-art prediction systems for pred icting socially plausible trajectories with relational reasoning. To validate th e ability of relational reasoning, we experiment with synthetic physics simulati ons to reflect the ability to capture group behaviors, reason interaction streng th and interaction category. To validate the effectiveness of prediction, we con duct extensive experiments on three real-world trajectory prediction datasets, i ncluding NBA, SDD and ETH-UCY; and we show that with GroupNet, the CVAE-based pr ediction system outperforms state-of-the-art methods. We also show that adding G roupNet will further improve the performance of previous state-of-the-art predic tion systems.
****

Unbiased Subclass Regularization for Semi-Supervised Semantic Segmentation
Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9968-9978

Semi-supervised semantic segmentation learns from small amounts of labelled images and large amounts of unlabelled images, which has witnessed impressive progress with the recent advance of deep neural networks. However, it often suffers from severe class-bias problem while exploring the unlabelled images, largely due to the clear pixel-wise class imbalance in the labelled images. This paper presents an unbiased subclass regularization network (USRN) that alleviates the class imbalance issue by learning class-unbiased segmentation from balanced subclass distributions. We build the balanced subclass distributions by clustering pixels of each original class into multiple subclasses of similar sizes, which provide class-balanced pseudo supervision to regularize the class-biased segmentation. In addition, we design an entropy-based gate mechanism to coordinate learning between the original classes and the clustered subclasses which facilitates subclass regularization effectively by suppressing unconfident subclass predictions. Extensive experiments over multiple public benchmarks show that USRN achieves superior performance as compared with the state-of-the-art. The code will be made available on Github.

****

P3IV: Probabilistic Procedure Planning From Instructional Videos With Weak Supervision

He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G. Derpanis, Richard P. Wildes, Allan D. Jepson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2938-2948

In this paper, we study the problem of procedure planning in instructional videos. Here, an agent must produce a plausible sequence of actions that can transform the environment from a given start to a desired goal state. When learning procedure planning from instructional videos, most recent work leverages intermediate visual observations as supervision, which requires expensive annotation efforts to localize precisely all the instructional steps in training videos. In contrast, we remove the need for expensive temporal video annotations and propose a weakly supervised approach by learning from natural language instructions. Our model is based on a transformer equipped with a memory module, which maps the start and goal observations to a sequence of plausible actions. Furthermore, we augment our model with a probabilistic generative module to capture the uncertainty inherent to procedure planning, an aspect largely overlooked by previous work. We evaluate our model on three datasets and show our weakly-supervised approach outperforms previous fully supervised state-of-the-art models on multiple metrics.

****

Hierarchical Nearest Neighbor Graph Embedding for Efficient Dimensionality Reduction

Saquib Sarfraz, Marios Koulakis, Constantin Seibold, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 336-345

Dimensionality reduction is crucial both for visualization and preprocessing high dimensional data for machine learning. We introduce a novel method based on a hierarchy built on 1-nearest neighbor graphs in the original space which is used to preserve the grouping properties of the data distribution on multiple levels. The core of the proposal is an optimization-free projection that is competitive with the latest versions of t-SNE and UMAP in performance and visualization quality while being an order of magnitude faster at run-time. Furthermore, its interpretable mechanics, the ability to project new data, and the natural separation of data clusters in visualizations make it a general purpose unsupervised dimension reduction technique. In the paper, we argue about the soundness of the proposed method and evaluate it on a diverse collection of datasets with sizes varying from 1K to 11M samples and dimensions from 28 to 16K. We perform comparisons with other state-of-the-art methods on multiple metrics and target dimensions highlighting its efficiency and performance. Code is available at https://github.com/koulakis/h-nne

****

Coupled Iterative Refinement for 6D Multi-Object Pose Estimation

Lahav Lipson, Zachary Teed, Ankit Goyal, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6728-6737
We address the task of 6D multi-object pose: given a set of known 3D objects and an RGB or RGB-D input image, we detect and estimate the 6D pose of each object. We propose a new approach to 6D object pose estimation which consists of an end-to-end differentiable architecture that makes use of geometric knowledge. Our approach iteratively refines both pose and correspondence in a tightly coupled manner, allowing us to dynamically remove outliers to improve accuracy. We use a novel differentiable layer to perform pose refinement by solving an optimization problem we refer to as Bidirectional Depth-Augmented Perspective-N-Point (BD-PnP). Our method achieves state-of-the-art accuracy on standard 6D Object Pose benchmarks.
****

Multi-View Transformer for 3D Visual Grounding
Shijia Huang, Yilun Chen, Jiaya Jia, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15524-15533
The 3D visual grounding task aims to ground a natural language description to the targeted object in a 3D scene, which is usually represented in 3D point clouds. Previous works studied visual grounding under specific views. The vision-language correspondence learned by this way can easily fail once the view changes. In this paper, we propose a Multi-View Transformer (MVT) for 3D visual grounding. We project the 3D scene to a multi-view space, in which the position information of the 3D scene under different views are modeled simultaneously and aggregated together. The multi-view space enables the network to learn a more robust multi-modal representation for 3D visual grounding and eliminates the dependence on specific views. Extensive experiments show that our approach significantly outperforms all state-of-the-art methods. Specifically, on Nr3D and Sr3D datasets, our method outperforms the best competitor by 11.2% and 7.1% and even surpasses recent work with extra 2D assistance by 5.9% and 6.6%. Our code is available at https://github.com/sega-hsj/MVT-3DVG.
****

Structured Sparse R-CNN for Direct Scene Graph Generation
Yao Teng, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19437-19446
Scene graph generation (SGG) is to detect object pairs with their relations in an image. Existing SGG approaches often use multi-stage pipelines to decompose this task into object detection, relation graph construction, and dense or dense-to-sparse relation prediction. Instead, from a perspective on SGG as a direct set prediction, this paper presents a simple, sparse, and unified framework, termed as Structured Sparse R-CNN. The key to our method is a set of learnable triplet queries and a structured triplet detector which could be jointly optimized from the training set in an end-to-end manner. Specifically, the triplet queries encode the general prior for object pairs with their relations, and provide an initial guess of scene graphs for subsequent refinement. The triplet detector presents a cascaded architecture to progressively refine the detected scene graphs with the customized dynamic heads. In addition, to relieve the training difficulty of our method, we propose a relaxed and enhanced training strategy based on knowledge distillation from a Siamese Sparse R-CNN. We perform experiments on several datasets: Visual Genome and Open Images V4/V6, and the results demonstrate that our method achieves the state-of-the-art performance. In addition, we also perform in-depth ablation studies to provide insights on our structured modeling in triplet detector design and training strategies. The code and models are made available at https://github.com/MCG-NJU/Structured-Sparse-RCNN.
****

Multi-Grained Spatio-Temporal Features Perceived Network for Event-Based Lip-Reading
Ganchao Tan, Yang Wang, Han Han, Yang Cao, Feng Wu, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20094-20103
Automatic lip-reading (ALR) aims to recognize words using visual information fro

m the speaker's lip movements. In this work, we introduce a novel type of sensin g device, event cameras, for the task of ALR. Event cameras have both technical and application advantages over conventional cameras for the ALR task because th ey have higher temporal resolution, less redundant visual information, and lower power consumption. To recognize words from the event data, we propose a novel M ulti-grained Spatio-Temporal Features Perceived Network (MSTP) to perceive fine-grained spatio-temporal features from microsecond time-resolved event data. Spec ifically, a multi-branch network architecture is designed, in which different gr ained spatio-temporal features are learned by operating at different frame rates . The branch operating on the low frame rate can perceive spatial complete but t emporal coarse features. While the branch operating on the high frame rate can p erceive spatial coarse but temporal refinement features. And a message flow modu le is devised to integrate the features from different branches, leading to perc eiving more discriminative spatio-temporal features. In addition, we present the first event-based lip-reading dataset (DVS-Lip) captured by the event camera. E xperimental results demonstrated the superiority of the proposed model compared to the state-of-the-art event-based action recognition models and video-based li p-reading models.

****

Semi-Supervised Video Paragraph Grounding With Contrastive Encoder

Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, Heng Tao Shen; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2466-2475

Video events grounding aims at retrieving the most relevant moments from an untr immed video in terms of a given natural language query. Most previous works focu s on Video Sentence Grounding (VSG), which localizes the moment with a sentence query. Recently, researchers extended this task to Video Paragraph Grounding (VP G) by retrieving multiple events with a paragraph. However, we find the existing VPG methods may not perform well on context modeling and highly rely on video-p aragraph annotations. To tackle this problem, we propose a novel VPG method term ed Semi-supervised Video-Paragraph TRansformer (SVPTR), which can more effective ly exploit contextual information in paragraphs and significantly reduce the dep endency on annotated data. Our SVPTR method consists of two key components: (1) a base model VPTR that learns the video-paragraph alignment with contrastive enc oders and tackles the lack of sentence-level contextual interactions and (2) a s emi-supervised learning framework with multimodal feature perturbations that red uces the requirements of annotated training data. We evaluate our model on three widely-used video grounding datasets, i.e., ActivityNet-Caption, Charades-CD-OO D, and TACoS. The experimental results show that our SVPTR method establishes th e new state-of-the-art performance on all datasets. Even under the conditions of fewer annotations, it can also achieve competitive results compared with recent VPG methods.

****

Continual Predictive Learning From Videos

Geng Chen, Wendong Zhang, Han Lu, Siyu Gao, Yunbo Wang, Mingsheng Long, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 10728-10737

Predictive learning ideally builds the world model of physical processes in one or more given environments. Typical setups assume that we can collect data from all environments at all times. In practice, however, different prediction tasks may arrive sequentially so that the environments may change persistently through out the training procedure. Can we develop predictive learning algorithms that c an deal with more realistic, non-stationary physical environments? In this paper , we study a new continual learning problem in the context of video prediction, and observe that most existing methods suffer from severe catastrophic forgettin g in this setup. To tackle this problem, we propose the continual predictive lea rning (CPL) approach, which learns a mixture world model via predictive experien ce replay and performs test-time adaptation with non-parametric task inference. We construct two new benchmarks based on RoboNet and KTH, in which different tas ks correspond to different physical robotic environments or human actions. Our a

pproach is shown to effectively mitigate forgetting and remarkably outperform th
e naive combinations of previous art in video prediction and continual learning.
****

Weakly Paired Associative Learning for Sound and Image Representations via Bimod
al Associative Memory
Sangmin Lee, Hyung-Il Kim, Yong Man Ro; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10534-10543
Data representation learning without labels has attracted increasing attention d
ue to its nature that does not require human annotation. Recently, representatio
n learning has been extended to bimodal data, especially sound and image which a
re closely related to basic human senses. Existing sound and image representatio
n learning methods necessarily require a large number of sound and image with co
rresponding pairs. Therefore, it is difficult to ensure the effectiveness of the
 methods in the weakly paired condition, which lacks paired bimodal data. In fac
t, according to human cognitive studies, the cognitive functions in the human br
ain for a certain modality can be enhanced by receiving other modalities, even n
ot directly paired ones. Based on the observation, we propose a new problem to d
eal with the weakly paired condition: How to boost a certain modal representatio
n even by using other unpaired modal data. To address the issue, we introduce a
novel bimodal associative memory (BMA-Memory) with key-value switching. It enabl
es to build sound-image association with small paired bimodal data and to boost
the built association with the easily obtainable large amount of unpaired data.
Through the proposed associative learning, it is possible to reinforce the repre
sentation of a certain modality (e.g., sound) even by using other unpaired modal
 data (e.g., images).
****

BARC: Learning To Regress 3D Dog Shape From Images by Exploiting Breed Informati
on
Nadine Rüegg, Silvia Zuffi, Konrad Schindler, Michael J. Black; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
pp. 3876-3884
Our goal is to recover the 3D shape and pose of dogs from a single image. This i
s a challenging task because dogs exhibit a wide range of shapes and appearances
, and are highly articulated. Recent work has proposed to directly regress the S
MAL animal model, with additional limb scale parameters, from images. Our method
, called BARC (Breed-Augmented Regression using Classification), goes beyond pri
or work in several important ways. First, we modify the SMAL shape space to be m
ore appropriate for representing dog shape. But, even with a better shape model,
 the problem of regressing dog shape from an image is still challenging because
we lack paired images with 3D ground truth. To compensate for the lack of paired
 data, we formulate novel losses that exploit information about dog breeds. In p
articular, we exploit the fact that dogs of the same breed have similar body sha
pes. We formulate a novel breed similarity loss consisting of two parts: One ter
m encourages the shape of dogs from the same breed to be more similar than dogs
of different breeds. The second one, a breed classification loss, helps to produ
ce recognizable breed-specific shapes. Through ablation studies, we find that ou
r breed losses significantly improve shape accuracy over a baseline without them
. We also compare BARC qualitatively to WLDO with a perceptual study and find th
at our approach produces dogs that are significantly more realistic. This work s
hows that a-priori information about genetic similarity can help to compensate f
or the lack of 3D training data. This concept may be applicable to other animal
species or groups of species. Our code is publicly available for research purpos
es at https://barc.is.tue.mpg.de/.
****

Knowledge Distillation: A Good Teacher Is Patient and Consistent
Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, Alexander
Kolesnikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2022, pp. 10925-10934
There is a growing discrepancy in computer vision between large-scale models tha
t achieve state-of-the-art performance and models that are affordable in practic

al applications. In this paper we address this issue and significantly bridge th e gap between these two types of models. Throughout our empirical investigation we do not aim to necessarily propose a new method, but strive to identify a robu st and effective recipe for making state-of-the-art large scale models affordabl e in practice. We demonstrate that, when performed correctly, knowledge distilla tion can be a powerful tool for reducing the size of large models without compro mising their performance. In particular, we uncover that there are certain impli cit design choices, which may drastically affect the effectiveness of distillati on. Our key contribution is the explicit identification of these design choices, which were not previously articulated in the literature. We back up our finding s by a comprehensive empirical study, demonstrate compelling results on a wide r ange of vision datasets and, in particular, obtain a state-of-the-art ResNet-50 model for ImageNet, which achieves 82.8% top-1 accuracy.

****

PCA-Based Knowledge Distillation Towards Lightweight and Content-Style Balanced Photorealistic Style Transfer Models

Tai-Yin Chiu, Danna Gurari; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 7844-7853

Photorealistic style transfer entails transferring the style of a reference imag e to another image so the result seems like a plausible photo. Our work is inspi red by the observation that existing models are slow due to their large sizes. W e introduce PCA-based knowledge distillation to distill lightweight models and s how it is motivated by theory. To our knowledge, this is the first knowledge dis tillation method for photorealistic style transfer. Our experiments demonstrate its versatility for use with different backbone architectures, VGG and MobileNet , across six image resolutions. Compared to existing models, our top-performing model runs at speeds 5-20x faster using at most 1% of the parameters. Additional ly, our distilled models achieve a better balance between stylization strength a nd content preservation than existing models. To support reproducing our method and models, we share the code at https://github.com/chiutaiyin/PCA-Knowledge-Dis tillation.

****

Frame Averaging for Equivariant Shape Space Learning

Matan Atzmon, Koki Nagano, Sanja Fidler, Sameh Khamis, Yaron Lipman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 022, pp. 631-641

The task of shape space learning involves mapping a train set of shapes to and f rom a latent representation space with good generalization properties. Often, re al-world collections of shapes have symmetries, which can be defined as transfor mations that do not change the essence of the shape. A natural way to incorporat e symmetries in shape space learning is to ask that the mapping to the shape spa ce (encoder) and mapping from the shape space (decoder) are equivariant to the r elevant symmetries. In this paper, we present a framework for incorporating equi variance in encoders and decoders by introducing two contributions: (i) adapting the recent Frame Averaging (FA) framework for building generic, efficient, and maximally expressive Equivariant autoencoders; and (ii) constructing autoencoder s equivariant to piecewise Euclidean motions applied to different parts of the s hape. To the best of our knowledge, this is the first fully piecewise Euclidean equivariant autoencoder construction. Training our framework is simple: it uses standard reconstruction losses, and does not require the introduction of new los ses. Our architectures are built of standard (backbone) architectures with the a ppropriate frame averaging to make them equivariant. Testing our framework on bo th rigid shapes dataset using implicit neural representations, and articulated s hape datasets using mesh-based neural networks show state-of-the-art generalizat ion to unseen test shapes, improving relevant baselines by a large margin. In pa rticular, our method demonstrates significant improvement in generalizing to uns een articulated poses.

****

Transformer Tracking With Cyclic Shifting Window Attention

Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang; Proceedings of the IEEE/C

VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8791-8800
Transformer architecture has been showing its great strength in visual object tracking, for its effective attention mechanism. Existing transformer-based approaches adopt the pixel-to-pixel attention strategy on flattened image features and unavoidably ignore the integrity of objects. In this paper, we propose a new transformer architecture with multi-scale cyclic shifting window attention for visual object tracking, elevating the attention from pixel to window level. The cross-window multi-scale attention has the advantage of aggregating attention at different scales and generates the best fine-scale match for the target object. Furthermore, the cyclic shifting strategy brings greater accuracy by expanding the window samples with positional information, and at the same time saves huge amounts of computational power by removing redundant calculations. Extensive experiments demonstrate the superior performance of our method, which also sets the new state-of-the-art records on five challenging datasets, along with the VOT2020, UAV123, LaSOT, TrackingNet, and GOT-10k benchmarks.
****

ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues
Hengcan Shi, Munawar Hayat, Yicheng Wu, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9611-9620
Object proposal generation is an important and fundamental task in computer vision. In this paper, we propose ProposalCLIP, a method towards unsupervised open-category object proposal generation. Unlike previous works which require a large number of bounding box annotations and/or can only generate proposals for limited object categories, our ProposalCLIP is able to predict proposals for a large variety of object categories without annotations, by exploiting CLIP (contrastive language-image pre-training) cues. Firstly, we analyze CLIP for unsupervised open-category proposal generation and design an objectness score based on our empirical analysis on proposal selection. Secondly, a graph-based merging module is proposed to solve the limitations of CLIP cues and merge fragmented proposals. Finally, we present a proposal regression module that extracts pseudo labels based on CLIP cues and trains a lightweight network to further refine proposals. Extensive experiments on PASCAL VOC, COCO and Visual Genome datasets show that our ProposalCLIP can better generate proposals than previous state-of-the-art methods. Our ProposalCLIP also shows benefits for downstream tasks, such as unsupervised object detection.
****

Towards Understanding Adversarial Robustness of Optical Flow Networks
Simon Schrodi, Tonmoy Saikia, Thomas Brox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8916-8924
Recent work demonstrated the lack of robustness of optical flow networks to physical patch-based adversarial attacks. The possibility to physically attack a basic component of automotive systems is a reason for serious concerns. In this paper, we analyze the cause of the problem and show that the lack of robustness is rooted in the classical aperture problem of optical flow estimation in combination with bad choices in the details of the network architecture. We show how these mistakes can be rectified in order to make optical flow networks robust to physical patch-based attacks. Additionally, we take a look at global white-box attacks in the scope of optical flow. We find that targeted white-box attacks can be crafted to bias flow estimation models towards any desired output, but this requires access to the input images and model weights. However, in the case of universal attacks, we find that optical flow networks are robust. Code is available at https://github.com/lmb-freiburg/understanding_flow_robustness.
****

Panoptic SegFormer: Delving Deeper Into Panoptic Segmentation With Transformers
Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, Tong Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1280-1289

Panoptic segmentation involves a combination of joint semantic segmentation and instance segmentation, where image contents are divided into two types: things a nd stuff. We present Panoptic SegFormer, a general framework for panoptic segmen tation with transformers. It contains three innovative components: an efficient deeply-supervised mask decoder, a query decoupling strategy, and an improved pos t-processing method. We also use Deformable DETR to efficiently process multi-sc ale features, which is a fast and efficient version of DETR. Specifically, we su pervise the attention modules in the mask decoder in a layer-wise manner. This d eep supervision strategy lets the attention modules quickly focus on meaningful semantic regions. It improves performance and reduces the number of required tra ining epochs by half compared to Deformable DETR. Our query decoupling strategy decouples the responsibilities of the query set and avoids mutual interference b etween things and stuff. In addition, our post-processing strategy improves perf ormance without additional costs by jointly considering classification and segme ntation qualities to resolve conflicting mask overlaps. Our approach increases t he accuracy 6.2% PQ over the baseline DETR model. Panoptic SegFormer achieves st ate-of-the-art results on COCO test-dev with 56.2% PQ. It also shows stronger ze ro-shot robustness over existing methods.

****

Training High-Performance Low-Latency Spiking Neural Networks by Differentiation on Spike Representation

Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, Zhi-Quan Luo; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12444-12453

Spiking Neural Network (SNN) is a promising energy-efficient AI model when imple mented on neuromorphic hardware. However, it is a challenge to efficiently train SNNs due to their non-differentiability. Most existing methods either suffer fr om high latency (i.e., long simulation time steps), or cannot achieve as high pe rformance as Artificial Neural Networks (ANNs). In this paper, we propose the Di fferentiation on Spike Representation (DSR) method, which could achieve high per formance that is competitive to ANNs yet with low latency. First, we encode the spike trains into spike representation using (weighted) firing rate coding. Base d on the spike representation, we systematically derive that the spiking dynamic s with common neural models can be represented as some sub-differentiable mappin g. With this viewpoint, our proposed DSR method trains SNNs through gradients of the mapping and avoids the common non-differentiability problem in SNN training . Then we analyze the error when representing the specific mapping with the forw ard computation of the SNN. To reduce such error, we propose to train the spike threshold in each layer, and to introduce a new hyperparameter for the neural mo dels. With these components, the DSR method can achieve state-of-the-art SNN per formance with low latency on both static and neuromorphic datasets, including CI FAR-10, CIFAR-100, ImageNet, and DVS-CIFAR10.

****

AnyFace: Free-Style Text-To-Face Synthesis and Manipulation

Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, Zhenan Sun; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 18687-18696

Existing text-to-image synthesis methods generally are only applicable to words in the training dataset. However, human faces are so variable to be described wi th limited words. So this paper proposes the first free-style text-to-face metho d namely AnyFace enabling much wider open world applications such as metaverse, social media, cosmetics, forensics, etc. AnyFace has a novel two-stream framewor k for face image synthesis and manipulation given arbitrary descriptions of the human face. Specifically, one stream performs text-to-face generation and the ot her conducts face image reconstruction. Facial text and image features are extra cted using the CLIP (Contrastive Language-Image Pre-training) encoders. And a co llaborative Cross Modal Distillation (CMD) module is designed to align the lingu istic and visual features across these two streams. Furthermore, a Diverse Tripl et Loss (DT loss) is developed to model fine-grained features and improve facial diversity. Extensive experiments on Multi-modal CelebA-HQ and CelebAText-HQ dem

onstrate significant advantages of AnyFace over state-of-the-art methods. AnyFac
e can achieve high-quality, high-resolution, and high-diversity face synthesis a
nd manipulation results without any constraints on the number and content of inp
ut captions.
****

HL-Net: Heterophily Learning Network for Scene Graph Generation
Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, Dacheng Tao; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 19476-19485
Scene graph generation (SGG) aims to detect objects and predict their pairwise r
elationships within an image. Current SGG methods typically utilize graph neural
 networks (GNNs) to acquire context information between objects/relationships. D
espite their effectiveness, however, current SGG methods only assume scene graph
 homophily while ignoring heterophily. Accordingly, in this paper, we propose a
novel Heterophily Learning Network (HL-Net) to comprehensively explore the homop
hily and heterophily between objects/relationships in scene graphs. More specifi
cally, HL-Net comprises the following 1) an adaptive reweighting transformer mod
ule, which adaptively integrates the information from different layers to exploi
t both the heterophily and homophily in objects; 2) a relationship feature propa
gation module that efficiently explores the connections between relationships by
 considering heterophily in order to refine the relationship representation; 3)
a heterophily-aware message-passing scheme to further distinguish the heterophil
y and homophily between objects/relationships, thereby facilitating improved mes
sage passing in graphs. We conducted extensive experiments on two public dataset
s: Visual Genome (VG) and Open Images (OI). The experimental results demonstrate
 the superiority of our proposed HL-Net over existing state-of-the-art approache
s. In more detail, HL-Net outperforms the second-best competitors by 2.1% on the
 VG dataset for scene graph classification and 1.2% on the IO dataset for the fi
nal score.
****

Lifelong Graph Learning
Chen Wang, Yuheng Qiu, Dasong Gao, Sebastian Scherer; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13719-
13728
Graph neural networks (GNN) are powerful models for many graph-structured tasks.
 Existing models often assume that the complete structure of the graph is availa
ble during training. In practice, however, graph-structured data is usually form
ed in a streaming fashion so that learning a graph continuously is often necessa
ry. In this paper, we bridge GNN and lifelong learning by converting a continual
 graph learning problem to a regular graph learning problem so GNN can inherit t
he lifelong learning techniques developed for convolutional neural networks (CNN
). We propose a new topology, the feature graph, which takes features as new nod
es and turns nodes into independent graphs. This successfully converts the origi
nal problem of node classification to graph classification. In the experiments,
we demonstrate the efficiency and effectiveness of feature graph networks (FGN)
by continuously learning a sequence of classical graph datasets. We also show th
at FGN achieves superior performance in two applications, i.e., lifelong human a
ction recognition with wearable devices and feature matching. To the best of our
 knowledge, FGN is the first method to bridge graph learning and lifelong learni
ng via a novel graph topology. Source code is available at https://github.com/wa
ng-chen/LGL
****

Hypergraph-Induced Semantic Tuplet Loss for Deep Metric Learning
Jongin Lim, Sangdoo Yun, Seulki Park, Jin Young Choi; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 212-22
2
In this paper, we propose Hypergraph-Induced Semantic Tuplet (HIST) loss for dee
p metric learning that leverages the multilateral semantic relations of multiple
 samples to multiple classes via hypergraph modeling. We formulate deep metric l
earning as a hypergraph node classification problem in which each sample in a mi

ni-batch is regarded as a node and each hyperedge models class-specific semantic relations represented by a semantic tuplet. Unlike previous graph-based losses that only use a bundle of pairwise relations, our HIST loss takes advantage of the multilateral semantic relations provided by the semantic tuplets through hypergraph modeling. Notably, by leveraging the rich multilateral semantic relations, HIST loss guides the embedding model to learn class-discriminative visual semantics, contributing to better generalization performance and model robustness against input corruptions. Extensive experiments and ablations provide a strong motivation for the proposed method and show that our HIST loss leads to improved feature learning, achieving state-of-the-art results on three widely used benchmarks. Code is available at https://github.com/ljin0429/HIST.

****

Computing Wasserstein-p Distance Between Images With Linear Cost

Yidong Chen, Chen Li, Zhonghua Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 519-528

When the images are formulated as discrete measures, computing Wasserstein-p distance between them is challenging due to the complexity of solving the corresponding Kantorovich's problem. In this paper, we propose a novel algorithm to compute the Wasserstein-p distance between discrete measures by restricting the optimal transport (OT) problem on a subset. First, we define the restricted OT problem and prove the solution of the restricted problem converges to antorovich's OT solution. Second, we propose the SparseSinkhorn algorithm for the restricted problem and provide a multi-scale algorithm to estimate the subset. Finally, we implement the proposed algorithm on CUDA and illustrate the linear computational cost in terms of time and memory requirements. We compute Wasserstein-p distance, estimate the transport mapping, and transfer color between color images with size ranges from 64x64 to 1920x1200. (Our code is available at https://github.com/ucascnic/CudaOT)

****

DLFormer: Discrete Latent Transformer for Video Inpainting

Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, Chen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3511-3520

Video inpainting remains a challenging problem to fill with plausible and coherent content in unknown areas in video frames despite the prevalence of data-driven methods. Although various transformer-based architectures yield promising result for this task, they still suffer from hallucinating blurry contents and long-term spatial-temporal inconsistency. While noticing the capability of discrete representation for complex reasoning and predictive learning, we propose a novel Discrete Latent Transformer (DLFormer) to reformulate video inpainting tasks into the discrete latent space rather the previous continuous feature space. Specifically, we first learn a unique compact discrete codebook and the corresponding autoencoder to represent the target video. Built upon these representative discrete codes obtained from the entire target video, the subsequent discrete latent transformer is capable to infer proper codes for unknown areas under a self-attention mechanism, and thus produces fine-grained content with long-term spatial-temporal consistency. Moreover, we further explicitly enforce the short-term consistency to relieve temporal visual jitters via a temporal aggregation block among adjacent frames. We conduct comprehensive quantitative and qualitative evaluations to demonstrate that our method significantly outperforms other state-of-the-art approaches in reconstructing visually-plausible and spatial-temporal coherent content with fine-grained details

****

Unsupervised Representation Learning for Binary Networks by Joint Classifier Learning

Dahyun Kim, Jonghyun Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9747-9756

Self-supervised learning is a promising unsupervised learning framework that has achieved success with large floating point networks. But such networks are not readily deployable to edge devices. To accelerate deployment of models with the

benefit of unsupervised representation learning to such resource limited devices for various downstream tasks, we propose a self-supervised learning method for binary networks that uses a moving target network. In particular, we propose to jointly train a randomly initialized classifier, attached to a pretrained floating point feature extractor, with a binary network. Additionally, we propose a feature similarity loss, a dynamic loss balancing and modified multi-stage training to further improve the accuracy, and call our method BURN. Our empirical validations over five downstream tasks using seven datasets show that BURN outperforms self-supervised baselines for binary networks and sometimes outperforms supervised pretraining. Code is availabe at https://github.com/naver-ai/burn.
****

High Quality Segmentation for Ultra High-Resolution Images
Tiancheng Shen, Yuechen Zhang, Lu Qi, Jason Kuen, Xingyu Xie, Jianlong Wu, Zhe Lin, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1310-1319
To segment 4K or 6K ultra high-resolution images needs extra computation consideration in image segmentation. Common strategies, such as down-sampling, patch cropping, and cascade model, cannot address well the balance issue between accuracy and computation cost. Motivated by the fact that humans distinguish among objects continuously from coarse to precise levels, we propose the Continuous Refinement Model(CRM) for the ultra high-resolution segmentation refinement task. CRM continuously aligns the feature map with the refinement target and aggregates features to reconstruct these images' details. Besides, our CRM shows its significant generalization ability to fill the resolution gap between low-resolution training images and ultra high-resolution testing ones. We present quantitative performance evaluation and visualization to show that our proposed method is fast and effective on image segmentation refinement.
****

Investigating Tradeoffs in Real-World Video Super-Resolution
Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5962-5971
The diversity and complexity of degradations in real-world video super-resolution (VSR) pose non-trivial challenges in inference and training. First, while long-term propagation leads to improved performance in cases of mild degradations, severe in-the-wild degradations could be exaggerated through propagation, impairing output quality. To balance the tradeoff between detail synthesis and artifact suppression, we found an image pre-cleaning stage indispensable to reduce noises and artifacts prior to propagation. Equipped with a carefully designed cleaning module, our RealBasicVSR outperforms existing methods in both quality and efficiency. Second, real-world VSR models are often trained with diverse degradations to improve generalizability, requiring increased batch size to produce a stable gradient. Inevitably, the increased computational burden results in various problems, including 1) speed-performance tradeoff and 2) batch-length tradeoff. To alleviate the first tradeoff, we propose a stochastic degradation scheme that reduces up to 40% of training time without sacrificing performance. We then analyze different training settings and suggest that employing longer sequences rather than larger batches during training allows more effective uses of temporal information, leading to more stable performance during inference. To facilitate fair comparisons, we propose the new VideoLQ dataset, which contains a large variety of real-world low-quality video sequences containing rich textures and patterns. Our dataset can serve as a common ground for benchmarking. Code, models, and the dataset will be made publicly available.
****

MERLOT Reserve: Neural Script Knowledge Through Vision and Language and Sound
Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, Yejin Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16375-16387
As humans, we navigate a multimodal world, building a holistic understanding fro

m all our senses. We introduce MERLOT Reserve, a model that represents videos jo intly over time -- through a new training objective that learns from audio, subt itles, and video frames. Given a video, we replace snippets of text and audio wi th a MASK token; the model learns by choosing the correct masked-out snippet. Ou r objective learns faster than alternatives, and performs well at scale: we pret rain on 20 million YouTube videos. Empirical results show that MERLOT Reserve le arns strong multimodal representations. When finetuned, it sets state-of-the-art on Visual Commonsense Reasoning (VCR), TVQA, and Kinetics-600; outperforming pr ior work by 5%, 7%, and 1.5% respectively. Ablations show that these tasks benef it from audio pretraining -- even VCR, a QA task centered around images (without sound). Moreover, our objective enables out-of-the-box prediction, revealing st rong multimodal commonsense understanding. In a fully zero-shot setting, our mod el obtains competitive results on four video tasks, even outperforming supervise d approaches on the recently proposed Situated Reasoning (STAR) benchmark. We an alyze why audio enables better vision-language representations, suggesting signi ficant opportunities for future research. We conclude by discussing ethical and societal implications of multimodal pretraining.

****

Differentiable Stereopsis: Meshes From Multiple Views Using Differentiable Rende ring

Shubham Goel, Georgia Gkioxari, Jitendra Malik; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8635-8644

We propose Differentiable Stereopsis, a multi-view stereo approach that reconstr ucts shape and texture from few input views and noisy cameras. We pair tradition al stereopsis and modern differentiable rendering to build an end-to-end model w hich predicts textured 3D meshes of objects with varying topologies and shape. W e frame stereopsis as an optimization problem and simultaneously update shape an d cameras via simple gradient descent. We run an extensive quantitative analysis and compare to traditional multi-view stereo techniques and state-of-the-art le arning based methods. We show compelling reconstructions on challenging real-wor ld scenes and for an abundance of object types with complex shape, topology and texture.

****

Towards Practical Certifiable Patch Defense With Vision Transformer

Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Wenqiang Zhang; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2022, pp. 15148-15158

Patch attacks, one of the most threatening forms of physical attack in adversari al examples, can lead networks to induce misclassification by modifying pixels a rbitrarily in a continuous region. Certifiable patch defense can guarantee robus tness that the classifier is not affected by patch attacks. Existing certifiable patch defenses sacrifice the clean accuracy of classifiers and only obtain a lo w certified accuracy on toy datasets. Furthermore, the clean and certified accur acy of these methods is still significantly lower than the accuracy of normal cl assification networks, which limits their application in practice. To move towar ds a practical certifiable patch defense, we introduce Vision Transformer (ViT) into the framework of Derandomized Smoothing (DS). Specifically, we propose a pr ogressive smoothed image modeling task to train Vision Transformer, which can ca pture the more discriminable local context of an image while preserving the glob al semantic information. For efficient inference and deployment in the real worl d, we innovatively reconstruct the global self-attention structure of the origin al ViT into isolated band unit self-attention. On ImageNet, under 2% area patch attacks our method achieves 41.70% certified accuracy, a nearly 1-fold increase over the previous best method (26.00%). Simultaneously, our method achieves 78.5 8% clean accuracy, which is quite close to the normal ResNet-101 accuracy. Exten sive experiments show that our method obtains state-of-the-art clean and certifi ed accuracy with inferring efficiently on CIFAR-10 and ImageNet.

****

A Conservative Approach for Unbiased Learning on Unknown Biases

Myeongho Jeon, Daekyung Kim, Woochul Lee, Myungjoo Kang, Joonseok Lee; Proceedin

gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16752-16760
Although convolutional neural networks (CNNs) achieve state-of-the-art in image classification, recent works address their unreliable predictions due to their excessive dependence on biased training data. Existing unbiased modeling postulates that the bias in the dataset is obvious to know, but it is actually unsuited for image datasets including countless sensory attributes. To mitigate this issue, we present a new scenario that does not necessitate a predefined bias. Under the observation that CNNs do have multi-variant and unbiased representations in the model, we propose a conservative framework that employs this internal information for unbiased learning. Specifically, this mechanism is implemented via hierarchical features captured along the multiple layers and orthogonal regularization. Extensive evaluations on public benchmarks demonstrate our method is effective for unbiased learning.
****

Large-Scale Video Panoptic Segmentation in the Wild: A Benchmark
Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21033-21043
In this paper, we present a new large-scale dataset for the video panoptic segmentation task, which aims to assign semantic classes and track identities to all pixels in a video. As the ground truth for this task is difficult to annotate, previous datasets for video panoptic segmentation are limited by either small scales or the number of scenes. In contrast, our large-scale VIdeo Panoptic Segmentation in the Wild (VIPSeg) dataset provides 3,536 videos and 84,750 frames with pixel-level panoptic annotations, covering a wide range of real-world scenarios and categories. To the best of our knowledge, our VIPSeg is the first attempt to tackle the challenging video panoptic segmentation task in the wild by considering diverse scenarios. Based on VIPSeg, we evaluate existing video panoptic segmentation approaches and propose an efficient and effective clip-based baseline method to analyze our VIPSeg dataset. Our dataset is available at https://github.com/VIPSeg-Dataset/VIPSeg-Dataset/.
****

Label, Verify, Correct: A Simple Few Shot Object Detection Method
Prannay Kaul, Weidi Xie, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14237-14247
The objective of this paper is few-shot object detection (FSOD) - the task of expanding an object detector for a new category given only a few instances as training. We introduce a simple pseudo-labelling method to source high-quality pseudo-annotations from the training set, for each new category, vastly increasing the number of training instances and reducing class imbalance; our method finds previously unlabelled instances. Naively training with model predictions yields sub-optimal performance; we present two novel methods to improve the precision of the pseudo-labelling process: first, we introduce a verification technique to remove candidate detections with incorrect class labels; second, we train a specialised model to correct poor quality bounding boxes. After these two novel steps, we obtain a large set of high-quality pseudo-annotations that allow our final detector to be trained end-to-end. Additionally, we demonstrate our method maintains base class performance, and the utility of simple augmentations in FSOD. While benchmarking on PASCAL VOC and MS-COCO, our method achieves state-of-the-art or second-best performance compared to existing approaches across all number of shots.
****

Aesthetic Text Logo Synthesis via Content-Aware Layout Inferring
Yizhi Wang, Guo Pu, Wenhan Luo, Yexin Wang, Pengfei Xiong, Hongwen Kang, Zhouhui Lian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2436-2445
Text logo design heavily relies on the creativity and expertise of professional designers, in which arranging element layouts is one of the most important procedures. However, few attention has been paid to this task which needs to take man

y factors (e.g., fonts, linguistics, topics, etc.) into consideration. In this p aper, we propose a content-aware layout generation network which takes glyph ima ges and their corresponding text as input and synthesizes aesthetic layouts for them automatically. Specifically, we develop a dual-discriminator module, includ ing a sequence discriminator and an image discriminator, to evaluate both the ch aracter placing trajectories and rendered shapes of synthesized text logos, resp ectively. Furthermore, we fuse the information of linguistics from texts and vis ual semantics from glyphs to guide layout prediction, which both play important roles in professional layout design. To train and evaluate our approach, we cons truct a dataset named as TextLogo3K, consisting of about 3,500 text logos and th eir pixel-level segmentation. Experimental studies on this dataset demonstrate t he effectiveness of our approach for synthesizing visually-pleasing text logos a nd verify its superiority against the state of the art.

****

Global Tracking via Ensemble of Local Trackers

Zikun Zhou, Jianqiu Chen, Wenjie Pei, Kaige Mao, Hongpeng Wang, Zhenyu He; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 8761-8770

The crux of long-term tracking lies in the difficulty of tracking the target wit h discontinuous moving caused by out-of-view or occlusion. Existing long-term tr acking methods follow two typical strategies. The first strategy employs a local tracker to perform smooth tracking and uses another re-detector to detect the t arget when the target is lost. While it can exploit the temporal context like hi storical appearances and locations of the target, a potential limitation of such strategy is that the local tracker tends to misidentify a nearby distractor as the target instead of activating the re-detector when the real target is out of view. The other long-term tracking strategy tracks the target in the entire imag e globally instead of local tracking based on the previous tracking results. Unf ortunately, such global tracking strategy cannot leverage the temporal context e ffectively. In this work, we combine the advantages of both strategies: tracking the target in a global view while exploiting the temporal context. Specifically , we perform global tracking via ensemble of local trackers spreading the full i mage. The smooth moving of the target can be handled steadily by one local track er. When the local tracker accidentally loses the target due to suddenly discont inuous moving, another local tracker close to the target is then activated and c an readily take over the tracking to locate the target. While the activated loca l tracker performs tracking locally by leveraging the temporal context, the ense mble of local trackers renders our model the global view for tracking. Extensive experiments on six datasets demonstrate that our method performs favorably agai nst state-of-the-art algorithms.

****

Autoregressive Image Generation Using Residual Quantization

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, Wook-Shin Han; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 11523-11532

For autoregressive (AR) modeling of high-resolution images, vector quantization (VQ) represents an image as a sequence of discrete codes. A short sequence lengt h is important for an AR model to reduce its computational costs to consider lon g-range interactions of codes. However, we postulate that previous VQ cannot sho rten the code sequence and generate high-fidelity images together in terms of th e rate-distortion trade-off. In this study, we propose the two-stage framework, which consists of Residual-Quantized VAE (RQ-VAE) and RQ-Transformer, to effecti vely generate high-resolution images. Given a fixed codebook size, RQ-VAE can pr ecisely approximate a feature map of an image and represent the image as a stack ed map of discrete codes. Then, RQ-Transformer learns to predict the quantized f eature vector at the next position by predicting the next stack of codes. Thanks to the precise approximation of RQ-VAE, we can represent a 256x256 image as 8x8 resolution of the feature map, and RQ-Transformer can efficiently reduce the co mputational costs. Consequently, our framework outperforms the existing AR model s on various benchmarks of unconditional and conditional image generation. Our a

pproach also has a significantly faster sampling speed than previous AR models t
o generate high-quality images.
****

MPC: Multi-View Probabilistic Clustering
Junjie Liu, Junlong Liu, Shaotian Yan, Rongxin Jiang, Xiang Tian, Boxuan Gu, Yao
wu Chen, Chen Shen, Jianqiang Huang; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2022, pp. 9509-9518
Despite the promising progress having been made, the two challenges of multi-vie
w clustering (MVC) are still waiting for better solutions: i) Most existing meth
ods are either not qualified or require additional steps for incomplete multi-vi
ew clustering and ii) noise or outliers might significantly degrade the overall
clustering performance. In this paper, we propose a novel unified framework for
incomplete and complete MVC named multi-view probabilistic clustering (MPC). MPC
 equivalently transforms multi-view pairwise posterior matching probability into
 composition of each view's individual distribution, which tolerates data missin
g and might extend to any number of views. Then graph-context-aware refinement w
ith path propagation and co-neighbor propagation is used to refine pairwise prob
ability, which alleviates the impact of noise and outliers. Finally, MPC also eq
uivalently transforms probabilistic clustering's objective to avoid complete pai
rwise computation and adjusts clustering assignments by maximizing joint probabi
lity iteratively. Extensive experiments on multiple benchmarks for incomplete an
d complete MVC show that MPC significantly outperforms previous state-of-the-art
 methods in both effectiveness and efficiency.
****

End-to-End Compressed Video Representation Learning for Generic Event Boundary D
etection
Congcong Li, Xinyao Wang, Longyin Wen, Dexiang Hong, Tiejian Luo, Libo Zhang; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2022, pp. 13967-13976
Generic event boundary detection aims to localize the generic, taxonomy-free eve
nt boundaries that segment videos into chunks. Existing methods typically requir
e video frames to be decoded before feeding into the network, which demands cons
iderable computational power and storage space. To that end, we propose a new en
d-to-end compressed video representation learning for event boundary detection t
hat leverages the rich information in the compressed domain, i.e., RGB, motion v
ectors, residuals, and the internal group of pictures (GOP) structure, without f
ully decoding the video. Specifically, we first use the ConvNets to extract feat
ures of the I-frames in the GOPs. After that, a light-weight spatial-channel com
pressed encoder is designed to compute the feature representations of the P-fram
es based on the motion vectors, residuals and representations of their dependent
 I-frames. A temporal contrastive module is proposed to determine the event boun
daries of video sequences. To remedy the ambiguities of annotations and speed up
 the training process, we use the Gaussian kernel to preprocess the ground-truth
 event boundaries. Extensive experiments conducted on the Kinetics-GEBD dataset
demonstrate that the proposed method achieves comparable results to the state-of
-the-art methods with 4.5x faster running speed.
****

GrainSpace: A Large-Scale Dataset for Fine-Grained and Domain-Adaptive Recogniti
on of Cereal Grains
Lei Fan, Yiwen Ding, Dongdong Fan, Donglin Di, Maurice Pagnucco, Yang Song; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 21116-21125
Cereal grains are a vital part of human diets and are important commodities for
people's livelihood and international trade. Grain Appearance Inspection (GAI) s
erves as one of the crucial steps for the determination of grain quality and gra
in stratification for proper circulation, storage and food processing, etc. GAI
is routinely performed manually by qualified inspectors with the aid of some han
d tools. Automated GAI has the benefit of greatly assisting inspectors with thei
r jobs but has been limited due to the lack of datasets and clear definitions of
 the tasks. In this paper we formulate GAI as three ubiquitous computer vision t

asks: fine-grained recognition, domain adaptation and out-of-distribution recogn
ition. We present a large-scale and publicly available cereal grains dataset cal
led GrainSpace. Specifically, we construct three types of device prototypes for
data acquisition, and a total of 5.25 million images determined by professional
inspectors. The grain samples including wheat, maize and rice are collected from
 five countries and more than 30 regions. We also develop a comprehensive benchm
ark based on semi-supervised learning and self-supervised learning techniques. T
o the best of our knowledge, GrainSpace is the first publicly released dataset f
or cereal grain inspection, https://github.com/hellodfan/GrainSpace.
****

BokehMe: When Neural Rendering Meets Classical Rendering
Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, Jianming Zhang; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 16283-16292
We propose BokehMe, a hybrid bokeh rendering framework that marries a neural ren
derer with a classical physically motivated renderer. Given a single image and a
 potentially imperfect disparity map, BokehMe generates high-resolution photo-re
alistic bokeh effects with adjustable blur size, focal plane, and aperture shape
. To this end, we analyze the errors from the classical scattering-based method
and derive a formulation to calculate an error map. Based on this formulation, w
e implement the classical renderer by a scattering-based method and propose a tw
o-stage neural renderer to fix the erroneous areas from the classical renderer.
The neural renderer employs a dynamic multi-scale scheme to efficiently handle a
rbitrary blur sizes, and it is trained to handle imperfect disparity input. Expe
riments show that our method compares favorably against previous methods on both
 synthetic image data and real image data with predicted disparity. A user study
 is further conducted to validate the advantage of our method.
****

Learning Modal-Invariant and Temporal-Memory for Video-Based Visible-Infrared Pe
rson Re-Identification
Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu
, David Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2022, pp. 20973-20982
Thanks for the cross-modal retrieval techniques, visible-infrared (RGB-IR) perso
n re-identification (Re-ID) is achieved by projecting them into a common space,
allowing person Re-ID in 24-hour surveillance systems. However, with respect to
the "probe-to-gallery", almost all existing RGB-IR based cross-modal person Re-I
D methods focus on image-to-image matching, while the video-to-video matching wh
ich contains much richer spatial- and temporal-information remains under-explore
d. In this paper, we primarily study the video-based cross-modal person Re-ID me
thod. To achieve this task, a video-based RGB-IR dataset is constructed, in whic
h 927 valid identities with 463,259 frames and 21,863 tracklets captured by 12 R
GB/IR cameras are collected. Based on our constructed dataset, we prove that wit
h the increase of frames in a tracklet, the performance does meet more enhanceme
nt, demonstrating the significance of video-to-video matching in RGB-IR person R
e-ID. Additionally, a novel method is further proposed, which not only projects
two modalities to a modal-invariant subspace, but also extracts the temporal-mem
ory for motion-invariant. Thanks to these two strategies, much better results ar
e achieved on our video-based cross-modal person Re-ID. The code is released at:
 https://github.com/VCM-project233/MITML.
****

MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning
Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian
Zhao, Xinge You; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2022, pp. 7612-7621
The key challenge of zero-shot learning (ZSL) is how to infer the latent semanti
c knowledge between visual and attribute features on seen classes, and thus achi
eving a desirable knowledge transfer to unseen classes. Prior works either simpl
y align the global features of an image with its associated class semantic vecto
r or utilize unidirectional attention to learn the limited latent semantic repre

sentations, which could not effectively discover the intrinsic semantic knowledg e (e.g., attribute semantics) between visual and attribute features. To solve th e above dilemma, we propose a Mutually Semantic Distillation Network (MSDN), whi ch progressively distills the intrinsic semantic representations between visual and attribute features for ZSL. MSDN incorporates an attribute->visual attention sub-net that learns attribute-based visual features, and a visual->attribute at tention sub-net that learns visual-based attribute features. By further introduc ing a semantic distillation loss, the two mutual attention sub-nets are capable of learning collaboratively and teaching each other throughout the training proc ess. The proposed MSDN yields significant improvements over the strong baselines , leading to new state-of-the-art performances on three popular challenging benc hmarks. Our source codes, pre-trained models, and more results have been availab le at the anonymous project website: https://anonymous.4open.science/r/MSDN.

****

Oriented RepPoints for Aerial Object Detection

Wentong Li, Yijie Chen, Kaixuan Hu, Jianke Zhu; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1829-1838

In contrast to the generic object, aerial targets are often non-axis aligned wit h arbitrary orientations having the cluttered surroundings. Unlike the mainstrea med approaches regressing the bounding box orientations, this paper proposes an effective adaptive points learning approach to aerial object detection by taking advantage of the adaptive points representation, which is able to capture the g eometric information of the arbitrary-oriented instances. To this end, three ori ented conversion functions are presented to facilitate the classification and lo calization with accurate orientation. Moreover, we propose an effective quality assessment and sample assignment scheme for adaptive points learning toward choo sing the representative oriented reppoints samples during training, which is abl e to capture the non-axis aligned features from adjacent objects or background n oises. A spatial constraint is introduced to penalize the outlier points for rou st adaptive learning. Experimental results on four challenging aerial datasets i ncluding DOTA, HRSC2016, UCAS-AOD and DIOR-R, demonstrate the efficacy of our pr oposed approach. The source code is availabel at: https://github.com/LiWentomng/ OrientedRepPoints.

****

OccAM's Laser: Occlusion-Based Attribution Maps for 3D Object Detectors on LiDAR Data

David Schinagl, Georg Krispel, Horst Possegger, Peter M. Roth, Horst Bischof; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1141-1150

While 3D object detection in LiDAR point clouds is well-established in academia and industry, the explainability of these models is a largely unexplored field. In this paper, we propose a method to generate attribution maps for the detected objects in order to better understand the behavior of such models. These maps i ndicate the importance of each 3D point in predicting the specific objects. Our method works with black-box models: We do not require any prior knowledge of the architecture nor access to the model's internals, like parameters, activations or gradients. Our efficient perturbation-based approach empirically estimates th e importance of each point by testing the model with randomly generated subsets of the input point cloud. Our sub-sampling strategy takes into account the speci al characteristics of LiDAR data, such as the depth-dependent point density. We show a detailed evaluation of the attribution maps and demonstrate that they are interpretable and highly informative. Furthermore, we compare the attribution m aps of recent 3D object detection architectures to provide insights into their d ecision-making processes.

****

BigDatasetGAN: Synthesizing ImageNet With Pixel-Wise Annotations

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, Antonio Torr alba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2022, pp. 21330-21340

Annotating images with pixel-wise labels is a time-consuming and costly process.

Recently, DatasetGAN showcased a promising alternative - to synthesize a large labeled dataset via a generative adversarial network (GAN) by exploiting a small set of manually labeled, GAN-generated images. Here, we scale DatasetGAN to ImageNet scale of class diversity. We take image samples from the class-conditional generative model BigGAN trained on ImageNet, and manually annotate only 5 images per class, for all 1k classes. By training an effective feature segmentation architecture on top of BigGAN, we turn BigGAN into a labeled dataset generator. We further show that VQGAN can similarly serve as a dataset generator, leveraging the already annotated data. We create a new ImageNet benchmark by labeling an additional set of real images and evaluate segmentation performance in a variety of settings. Through an extensive ablation study we show big gains in leveraging a large generated dataset to train different supervised and self-supervised backbone models on pixel-wise tasks. Furthermore, we demonstrate that using our synthesized datasets for pre-training leads to improvements over standard ImageNet pre-training on several downstream datasets, such as PASCAL-VOC, MS-COCO, Cityscapes and chest X-ray, as well as tasks (detection, segmentation). Our benchmark will be made public and maintain a leaderboard for this challenging task.
****

Align Representations With Base: A New Approach to Self-Supervised Learning
Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16600-16609
Existing symmetric contrastive learning methods suffer from collapses (complete and dimensional) or quadratic complexity of objectives. Departure from these methods which maximize mutual information of two generated views, along either instance or feature dimension, the proposed paradigm introduces intermediate variables at the feature level, and maximizes the consistency between variables and representations of each view. Specifically, the proposed intermediate variables are the nearest group of base vectors to representations. Hence, we call the proposed method ARB (Align Representations with Base). Compared with other symmetric approaches, ARB 1) does not require negative pairs, which leads the complexity of the overall objective function is in linear order, 2) reduces feature redundancy, increasing the information density of training samples, 3) is more robust to output dimension size, which outperforms previous feature-wise arts over 28% Top-1 accuracy on ImageNet-100 under low-dimension settings.
****

Exploring Denoised Cross-Video Contrast for Weakly-Supervised Temporal Action Localization
Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, Li Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19914-19924
Weakly-supervised temporal action localization aims to localize actions in untrimmed videos with only video-level labels. Most existing methods address this problem with a "localization-by-classification" pipeline that localizes action regions based on snippet-wise classification sequences. Snippet-wise classifications are unfortunately error prone due to the sparsity of video-level labels. Inspired by recent success in unsupervised contrastive representation learning, we propose a novel denoised cross-video contrastive algorithm, aiming to enhance the feature discrimination ability of video snippets for accurate temporal action localization in the weakly-supervised setting. This is enabled by three key designs: 1) an effective pseudo-label denoising module to alleviate the side effects caused by noisy contrastive features, 2) an efficient region-level feature contrast strategy with a region-level memory bank to capture "global" contrast across the entire dataset, and 3) a diverse contrastive learning strategy to enable action-background separation as well as intra-class compactness & inter-class separability. Extensive experiments on THUMOS14 and ActivityNet v1.3 demonstrate the superior performance of our approach.
****

Pre-Train, Self-Train, Distill: A Simple Recipe for Supersizing 3D Reconstruction

Kalyan Vasudev Alwala, Abhinav Gupta, Shubham Tulsiani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3773-3782

Our work learns a unified model for single-view 3D reconstruction of objects from hundreds of semantic categories. As a scalable alternative to direct 3D supervision, our work relies on segmented image collections for learning 3D of generic categories. Unlike prior works that use similar supervision but learn independent category-specific models from scratch, our approach of learning a unified model simplifies the training process while also allowing the model to benefit from the common structure across categories. Using image collections from standard recognition datasets, we show that our approach allows learning 3D inference for over 150 object categories. We evaluate using two datasets and qualitatively and quantitatively show that our unified reconstruction approach improves over prior category-specific reconstruction baselines. Our final 3D reconstruction model is also capable of zero-shot inference on images from unseen object categories and we empirically show that increasing the number of training categories improves the reconstruction quality.

****

Meta Distribution Alignment for Generalizable Person Re-Identification

Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2487-2496

Domain Generalizable (DG) person ReID is a challenging task which trains a model on source domains yet generalizes well on target domains. Existing methods use source domains to learn domain-invariant features, and assume those features are also irrelevant with target domains. However, they do not consider the target domain information which is unavailable in the training phrase of DG. To address this issue, we propose a novel Meta Distribution Alignment (MDA) method to enable them to share similar distribution in a test-time-training fashion. Specifically, since high-dimensional features are difficult to constrain with a known simple distribution, we first introduce an intermediate latent space constrained to a known prior distribution. The source domain data is mapped to this latent space and then reconstructed back. A meta-learning strategy is introduced to facilitate generalization and support fast adaption. To reduce their discrepancy, we further propose a test-time adaptive updating strategy based on the latent space which efficiently adapts model to unseen domains with a few samples. Extensive experimental results show that our model outperforms the state-of-the-art methods by up to 5.1% R-1 on average on the large-scale and 4.7% R-1 on the single-source domain generalization ReID benchmark.

****

TeachAugment: Data Augmentation Optimization Using Teacher Knowledge

Teppei Suzuki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10904-10914

Optimization of image transformation functions for the purpose of data augmentation has been intensively studied. In particular, adversarial data augmentation strategies, which search augmentation maximizing task loss, show significant improvement in the model generalization for many tasks. However, the existing methods require careful parameter tuning to avoid excessively strong deformations that take away image features critical for acquiring generalization. In this paper, we propose a data augmentation optimization method based on the adversarial strategy called TeachAugment, which can produce informative transformed images to the model without requiring careful tuning by leveraging a teacher model. Specifically, the augmentation is searched so that augmented images are adversarial for the target model and recognizable for the teacher model. We also propose data augmentation using neural networks, which simplifies the search space design and allows for updating of the data augmentation using the gradient method. We show that TeachAugment outperforms existing methods in experiments of image classification, semantic segmentation, and unsupervised representation learning tasks.

****

SVIP: Sequence VerIfication for Procedures in Videos

Yicheng Qian, Weixin Luo, Dongze Lian, Xu Tang, Peilin Zhao, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19890-19902

In this paper, we propose a novel sequence verification task that aims to distinguish positive video pairs performing the same action sequence from negative ones with step-level transformations but still conducting the same task. Such a challenging task resides in an open-set setting without prior action detection or segmentation that requires event-level or even frame-level annotations. To that end, we carefully reorganize two publicly available action-related datasets with step-procedure-task structure. To fully investigate the effectiveness of any method, we collect a scripted video dataset enumerating all kinds of step-level transformations in chemical experiments. Besides, a novel evaluation metric Weighted Distance Ratio is introduced to ensure equivalence for different step-level transformations during evaluation. In the end, a simple but effective baseline based on the transformer encoder with a novel sequence alignment loss is introduced to better characterize long-term dependency between steps, which outperforms other action recognition methods. Codes and data will be released.

****

Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning

Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15555-15564

Temporal sentence grounding aims to detect the most salient moment corresponding to the natural language query from untrimmed videos. As labeling the temporal boundaries is labor-intensive and subjective, the weakly-supervised methods have recently received increasing attention. Most of the existing weakly-supervised methods generate the proposals by sliding windows, which are content-independent and of low quality. Moreover, they train their model to distinguish positive visual-language pairs from negative ones randomly collected from other videos, ignoring the highly confusing video segments within the same video. In this paper, we propose Contrastive Proposal Learning (CPL) to overcome the above limitations. Specifically, we use multiple learnable Gaussian functions to generate both positive and negative proposals within the same video that can characterize the multiple events in a long video. Then, we propose a controllable easy to hard negative proposal mining strategy to collect negative samples within the same video, which can ease the model optimization and enables CPL to distinguish highly confusing scenes. The experiments show that our method achieves state-of-the-art performance on Charades-STA and ActivityNet Captions datasets. The code and models are available at https://github.com/minghangz/cpl.

****

Low-Resource Adaptation for Personalized Co-Speech Gesture Generation

Chaitanya Ahuja, Dong Won Lee, Louis-Philippe Morency; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20566-20576

Personalizing an avatar for co-speech gesture generation from spoken language requires learning the idiosyncrasies of a person's gesture style from a small amount of data. Previous methods in gesture generation require large amounts of data for each speaker, which is often infeasible. We propose an approach, named DiffGAN, that efficiently personalizes co-speech gesture generation models of a high-resource source speaker to target speaker with just 2 minutes of target training data. A unique characteristic of DiffGAN is its ability to account for the crossmodal grounding shift, while also addressing the distribution shift in the output domain. We substantiate the effectiveness of our approach a large scale publicly available dataset through quantitative, qualitative and user studies, which show that our proposed methodology significantly outperforms prior approaches for low-resource adaptation of gesture generation. Code and videos can be found at https://chahuja.com/diffgan

****

BoosterNet: Improving Domain Generalization of Deep Neural Nets Using Culpabilit

y-Ranked Features
Nourhan Bayasi, Ghassan Hamarneh, Rafeef Garbi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 538-548
Deep learning (DL) models trained to minimize empirical risk on a single domain often fail to generalize when applied to other domains. Model failures due to poor generalizability are quite common in practice and may prove quite perilous in mission-critical applications, e.g., diagnostic imaging where real-world data often exhibits pronounced variability. Such limitations have led to increased interest in domain generalization (DG) approaches that improve the ability of models learned from a single or multiple source domains to generalize to out-of-distribution (OOD) test domains. In this work, we propose BoosterNet, a lean add-on network that can be simply appended to any arbitrary core network to improve its generalization capability without requiring any changes in its architecture or training procedure. Specifically, using a novel measure of feature culpability, BoosterNet is trained episodically on the most and least culpable data features extracted from critical units in the core network based on their contribution towards class-specific prediction errors, which have shown to improve generalization. At inference time, corresponding test image features are extracted from the closest class-specific units, determined by smart gating via a Siamese network, and fed to BoosterNet for improved generalization. We evaluate the performance of BoosterNet within two very different classification problems, digits and skin lesions, and demonstrate a marked improvement in model generalization to OOD test domains compared to SOTA.
****

Task-Specific Inconsistency Alignment for Domain Adaptive Object Detection
Liang Zhao, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14217-14226
Detectors trained with massive labeled data often exhibit dramatic performance degradation in some particular scenarios with data distribution gap. To alleviate this problem of domain shift, conventional wisdom typically concentrates solely on reducing the discrepancy between the source and target domains via attached domain classifiers, yet ignoring the difficulty of such transferable features in coping with both classification and localization subtasks in object detection. To address this issue, in this paper, we propose Task-specific Inconsistency Alignment (TIA), by developing a new alignment mechanism in separate task spaces, improving the performance of the detector on both subtasks. Specifically, we add a set of auxiliary predictors for both classification and localization branches, and exploit their behavioral inconsistencies as finer-grained domain-specific measures. Then, we devise task-specific losses to align such cross-domain disagreement of both subtasks. By optimizing them individually, we are able to well approximate the category- and boundary-wise discrepancies in each task space, and therefore narrow them in a decoupled manner. TIA demonstrates superior results on various scenarios to the previous state-of-the-art methods. It is also observed that both the classification and localization capabilities of the detector are sufficiently strengthened, further demonstrating the effectiveness of our TIA method. Code and trained models are publicly available at https://github.com/MCG-NJU/TIA.
****

HDR-NeRF: High Dynamic Range Neural Radiance Fields
Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, Qing Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18398-18408
We present High Dynamic Range Neural Radiance Fields (HDR-NeRF) to recover an HDR radiance field from a set of low dynamic range (LDR) views with different exposures. Using the HDR-NeRF, we are able to generate both novel HDR views and novel LDR views under different exposures. The key to our method is to model the physical imaging process, which dictates that the radiance of a scene point transforms to a pixel value in the LDR image with two implicit functions: a radiance field and a tone mapper. The radiance field encodes the scene radiance (values vary from 0 to +infty), which outputs the density and radiance of a ray by giving c

orresponding ray origin and ray direction. The tone mapper models the mapping pr ocess that a ray hitting on the camera sensor becomes a pixel value. The color o f the ray is predicted by feeding the radiance and the corresponding exposure ti me into the tone mapper. We use the classic volume rendering technique to projec t the output radiance, colors, and densities into HDR and LDR images, while only the input LDR images are used as the supervision. We collect a new forward-faci ng HDR dataset to evaluate the proposed method. Experimental results on syntheti c and real-world scenes validate that our method can not only accurately control the exposures of synthesized views but also render views with a high dynamic ra nge.
****

MS2DG-Net: Progressive Correspondence Learning via Multiple Sparse Semantics Dyn amic Graph
Luanyuan Dai, Yizhang Liu, Jiayi Ma, Lifang Wei, Taotao Lai, Changcai Yang, Riqi ng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 8973-8982
Establishing superior-quality correspondences in an image pair is pivotal to man y subsequent computer vision tasks. Using Euclidean distance between corresponde nces to find neighbors and extract local information is a common strategy in pre vious works. However, most such works ignore similar sparse semantics informatio n between two given images and cannot capture local topology among correspondenc es well. Therefore, to deal with the above problems, Multiple Sparse Semantics D ynamic Graph Network (MS^ 2 DG-Net) is proposed, in this paper, to predict proba bilities of correspondences as inliers and recover camera poses. MS^ 2 DG-Net dy namically builds sparse semantics graphs based on sparse semantics similarity be tween two given images, to capture local topology among correspondences, while m aintaining permutation-equivariant. Extensive experiments prove that MS^ 2 DG-Ne t outperforms state-of-the-art methods in outlier removal and camera pose estima tion tasks on the public datasets with heavy outliers. Source code:https://githu b.com/changcaiyang/MS2DG-Net
****

Neural Emotion Director: Speech-Preserving Semantic Control of Facial Expression s in "In-the-Wild" Videos
Foivos Paraperas Papantoniou, Panagiotis P. Filntisis, Petros Maragos, Anastasio s Roussos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18781-18790
In this paper, we introduce a novel deep learning method for photo-realistic man ipulation of the emotional state of actors in "in-the-wild" videos. The proposed method is based on a parametric 3D face representation of the actor in the inpu t scene that offers a reliable disentanglement of the facial identity from the h ead pose and facial expressions. It then uses a novel deep domain translation fr amework that alters the facial expressions in a consistent and plausible manner, taking into account their dynamics. Finally, the altered facial expressions are used to photo-realistically manipulate the facial region in the input scene bas ed on an especially-designed neural face renderer. To the best of our knowledge, our method is the first to be capable of controlling the actor's facial express ions by even using as a sole input the semantic labels of the manipulated emotio ns, while at the same time preserving the speech-related lip movements. We condu ct extensive qualitative and quantitative evaluations and comparisons, which dem onstrate the effectiveness of our approach and the especially promising results that we obtain. Our method opens a plethora of new possibilities for useful appl ications of neural rendering technologies, ranging from movie post-production an d video games to photo-realistic affective avatars.
****

Learning To Listen: Modeling Non-Deterministic Dyadic Facial Motion
Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, Shiry Ginosar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20395-20405
We present a framework for modeling interactional communication in dyadic conver sations: given multimodal inputs of a speaker, we autoregressively output multip

le possibilities of corresponding listener motion. We combine the motion and spe ech audio of the speaker using a motion-audio cross attention transformer. Furth ermore, we enable non-deterministic prediction by learning a discrete latent rep resentation of realistic listener motion with a novel motion-encoding VQ-VAE. Ou r method organically captures the multimodal and non-deterministic nature of non verbal dyadic interactions. Moreover, it produces realistic 3D listener facial m otion synchronous with the speaker (see video). We demonstrate that our method o utperforms baselines qualitatively and quantitatively via a rich suite of experi ments. To facilitate this line of research, we introduce a novel and large in-th e-wild dataset of dyadic conversations. Code, data, and videos available at http ://people.eecs.berkeley.edu/ evonne_ng/projects/learning2listen/.
****

3PSDF: Three-Pole Signed Distance Function for Learning Surfaces With Arbitrary Topologies

Weikai Chen, Cheng Lin, Weiyang Li, Bo Yang; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18522-18531

Recent advances in learning 3D shapes using neural implicit functions have achie ved impressive results by breaking the previous barrier of resolution and divers ity for varying topologies. However, most of such approaches are limited to clos ed surfaces as they require the space to be divided into inside and outside. Mor e recent works based on unsigned distance function have been proposed to handle complex geometry containing both the open and closed surfaces. Nonetheless, as t heir direct outputs are point clouds, robustly obtaining high-quality meshing re sults from discrete points remains an open question. We present a novel learnabl e implicit representation, called the three-pole signed distance function (3PSDF), that can represent non-watertight 3D shapes with arbitrary topologies while s upporting easy field-to-mesh conversion using the classic Marching Cubes algorit hm. The key to our method is the introduction of a new sign, the NULL sign, in a ddition to the conventional in and out labels. The existence of the null sign co uld stop the formation of a closed isosurface derived from the bisector of the i n/out regions. Further, we propose a dedicated learning framework to effectively learn 3PSDF without worrying about the vanishing gradient due to the null label s. Experimental results show that our approach outperforms the previous state-of -the-art methods in a wide range of benchmarks both quantitatively and qualitati vely.
****

Capturing Humans in Motion: Temporal-Attentive 3D Human Pose and Shape Estimatio n From Monocular Video

Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, Hong-Yuan Mark Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13211-13220

Learning to capture human motion is essential to 3D human pose and shape estimat ion from monocular video. However, the existing methods mainly rely on recurrent or convolutional operation to model such temporal information, which limits the ability to capture non-local context relations of human motion. To address this problem, we propose a motion pose and shape network (MPS-Net) to effectively ca pture humans in motion to estimate accurate and temporally coherent 3D human pos e and shape from a video. Specifically, we first propose a motion continuity att ention (MoCA) module that leverages visual cues observed from human motion to ad aptively recalibrate the range that needs attention in the sequence to better ca pture the motion continuity dependencies. Then, we develop a hierarchical attent ive feature integration (HAFI) module to effectively combine adjacent past and f uture feature representations to strengthen temporal correlation and refine the feature representation of the current frame. By coupling the MoCA and HAFI modul es, the proposed MPS-Net excels in estimating 3D human pose and shape in the vid eo. Though conceptually simple, our MPS-Net not only outperforms the state-of-th e-art methods on the 3DPW, MPI-INF-3DHP, and Human3.6M benchmark datasets, but a lso uses fewer network parameters. The video demos can be found at https://mps-n et.github.io/MPS-Net/.
****

MixFormer: End-to-End Tracking With Iterative Mixed Attention
Yutao Cui, Cheng Jiang, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13608-13618
Tracking often uses a multi-stage pipeline of feature extraction, target information integration, and bounding box estimation. To simplify this pipeline and unify the process of feature extraction and target information integration, we present a compact tracking framework, termed as MixFormer, built upon transformers. Our core design is to utilize the flexibility of attention operations, and propose a Mixed Attention Module (MAM) for simultaneous feature extraction and target information integration. This synchronous modeling scheme allows to extract target-specific discriminative features and perform extensive communication between target and search area. Based on MAM, we build our MixFormer tracking framework simply by stacking multiple MAMs with progressive patch embedding and placing a localization head on top. In addition, to handle multiple target templates during online tracking, we devise an asymmetric attention scheme in MAM to reduce computational cost, and propose an effective score prediction module to select high-quality templates. Our MixFormer sets a new state-of-the-art performance on five tracking benchmarks, including LaSOT, TrackingNet, VOT2020, GOT-10k, and UAV123. In particular, our MixFormer-L achieves NP score of 79.9% on LaSOT, 88.9% on TrackingNet and EAO of 0.555 on VOT2020. We also perform in-depth ablation studies to demonstrate the effectiveness of simultaneous feature extraction and information integration. Code and trained models are publicly available at https://github.com/MCG-NJU/MixFormer.
****

Sparse Fuse Dense: Towards High Quality 3D Detection With Depth Completion
Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, Deng Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5418-5427
Current LiDAR-only 3D detection methods inevitably suffer from the sparsity of point clouds. Many multi-modal methods are proposed to alleviate this issue, while different representations of images and point clouds make it difficult to fuse them, resulting in suboptimal performance. In this paper, we present a novel multi-modal framework SFD (Sparse Fuse Dense), which utilizes pseudo point clouds generated from depth completion to tackle the issues mentioned above. Different from prior works, we propose a new RoI fusion strategy 3D-GAF (3D Grid-wise Attentive Fusion) to make fuller use of information from different types of point clouds. Specifically, 3D-GAF fuses 3D RoI features from the pair of point clouds in a grid-wise attentive way, which is more fine-grained and more precise. In addition, we propose a SynAugment (Synchronized Augmentation) to enable our multi-modal framework to utilize all data augmentation approaches tailored to LiDAR-only methods. Lastly, we customize an effective and efficient feature extractor CPConv (Color Point Convolution) for pseudo point clouds. It can explore 2D image features and 3D geometric features of pseudo point clouds simultaneously. Our method holds the highest entry on the KITTI car 3D object detection leaderboard, demonstrating the effectiveness of our SFD. Code will be made publicly available.
****

GIRAFFE HD: A High-Resolution 3D-Aware Generative Model
Yang Xue, Yuheng Li, Krishna Kumar Singh, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18440-18449
3D-aware generative models have shown that the introduction of 3D information can lead to more controllable image generation. In particular, the current state-of-the-art model GIRAFFE can control each object's rotation, translation, scale, and scene camera pose without corresponding supervision. However, GIRAFFE only operates well when the image resolution is low. We propose GIRAFFE HD, a high-resolution 3D-aware generative model that inherits all of GIRAFFE's controllable features while generating high-quality, high-resolution images (512^2 resolution and above). The key idea is to leverage a style-based neural renderer, and to independently generate the foreground and background to force their disentanglement while imposing consistency constraints to stitch them together to composite a c

oherent final image. We demonstrate state-of-the-art 3D controllable high-resolu tion image generation on multiple natural image datasets.
****

InOut: Diverse Image Outpainting via GAN Inversion
Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2022, pp. 11431-11440
Image outpainting seeks for a semantically consistent extension of the input ima ge beyond its available content. Compared to inpainting --- filling in missing p ixels in a way coherent with the neighboring pixels --- outpainting can be achie ved in more diverse ways since the problem is less constrained by the surroundin g pixels. Existing image outpainting methods pose the problem as a conditional i mage-to-image translation task, often generating repetitive structures and textu res by replicating the content available in the input image. In this work, we fo rmulate the problem from the perspective of inverting generative adversarial net works. Our generator renders micro-patches conditioned on their joint latent cod e as well as their individual positions in the image. To outpaint an image, we s eek for multiple latent codes not only recovering available patches but also syn thesizing diverse outpainting by patch-based generation. This leads to richer st ructure and content in the outpainted regions. Furthermore, our formulation allo ws for outpainting conditioned on the categorical input, thereby enabling flexib le user controls. Extensive experimental results demonstrate the proposed method performs favorably against existing in- and outpainting methods, featuring high er visual quality and diversity.
****

PNP: Robust Learning From Noisy Labels by Probabilistic Noise Prediction
Zeren Sun, Fumin Shen, Dan Huang, Qiong Wang, Xiangbo Shu, Yazhou Yao, Jinhui Ta ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 5311-5320
Label noise has been a practical challenge in deep learning due to the strong ca pability of deep neural networks in fitting all training data. Prior literature primarily resorts to sample selection methods for combating noisy labels. Howeve r, these approaches focus on dividing samples by order sorting or threshold sele ction, inevitably introducing hyper-parameters (e.g., selection ratio / threshol d) that are hard-to-tune and dataset-dependent. To this end, we propose a simple yet effective approach named PNP (Probabilistic Noise Prediction) to explicitly model label noise. Specifically, we simultaneously train two networks, in which one predicts the category label and the other predicts the noise type. By predi cting label noise probabilistically, we identify noisy samples and adopt dedicat ed optimization objectives accordingly. Finally, we establish a joint loss for n etwork update by unifying the classification loss, the auxiliary constraint loss , and the in-distribution consistency loss. Comprehensive experimental results o n synthetic and real-world datasets demonstrate the superiority of our proposed method.
****

Estimating Structural Disparities for Face Models
Shervin Ardeshir, Cristina Segalin, Nathan Kallus; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10358-103 67
In machine learning, disparity metrics are often defined by measuring the differ ence in the performance or outcome of a model, across different sub-populations (groups) of datapoints. Thus, the inputs to disparity quantification consist of a model's predictions y', the ground-truth labels for the predictions y, and gro up labels g for the data points. Performance of the model for each group is calc ulated by comparing y' and y for the datapoints within a specific group, and as a result, disparity of performance across the different groups can be calculated . In many real world scenarios however, group labels (g) may not be available at scale during training and validation time, or collecting them might not be feas ible or desirable as they could often be sensitive information. As a result, eva luating disparity metrics across categorical groups would not be feasible. On th

e other hand, in many scenarios noisy groupings may be obtainable using some for m of a proxy, which would allow measuring disparity metrics across sub-populatio ns. Here we explore performing such analysis on computer vision models trained o n human faces, and on tasks such as face attribute prediction and affect estimat ion. Our experiments indicate that embeddings resulting from an off-the-shelf fa ce recognition model, could meaningfully serve as a proxy for such estimation.
****

Revisiting the Transferability of Supervised Pretraining: An MLP Perspective
Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, Wanli Ouya ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 9183-9193
The pretrain-finetune paradigm is a classical pipeline in visual learning. Recen t progress on unsupervised pretraining methods shows superior transfer performan ce to their supervised counterparts. This paper revisits this phenomenon and she ds new light on understanding the transferability gap between unsupervised and s upervised pretraining from a multilayer perceptron (MLP) perspective. While prev ious works focus on the effectiveness of MLP on unsupervised image classificatio n where pretraining and evaluation are conducted on the same dataset, we reveal that the MLP projector is also the key factor to better transferability of unsup ervised pretraining methods than supervised pretraining methods. Based on this o bservation, we attempt to close the transferability gap between supervised and u nsupervised pretraining by adding an MLP projector before the classifier in supe rvised pretraining. Our analysis indicates that the MLP projector can help retai n intra-class variation of visual features, decrease the feature distribution di stance between pretraining and evaluation datasets, and reduce feature redundanc y. Extensive experiments on public benchmarks demonstrate that the added MLP pro jector significantly boosts the transferability of supervised pretraining, e.g. +7.2% top-1 accuracy on the concept generalization task, +5.8% top-1 accuracy fo r linear evaluation on 12-domain classification tasks, and +0.8% AP on COCO obje ct detection task, making supervised pretraining comparable or even better than unsupervised pretraining.
****

Plenoxels: Radiance Fields Without Neural Networks
Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, Angj oo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 5501-5510
We introduce Plenoxels (plenoptic voxels), a system for photorealistic view synt hesis. Plenoxels represent a scene as a sparse 3D grid with spherical harmonics. This representation can be optimized from calibrated images via gradient method s and regularization without any neural components. On standard, benchmark tasks , Plenoxels are optimized two orders of magnitude faster than Neural Radiance Fi elds with no loss in visual quality. For video and code, please see https://alex yu.net/plenoxels.
****

What Matters for Meta-Learning Vision Regression Tasks?
Ning Gao, Hanna Ziesche, Ngo Anh Vien, Michael Volpp, Gerhard Neumann; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14776-14786
Meta-learning is widely used in few-shot classification and function regression due to its ability to quickly adapt to unseen tasks. However, it has not yet bee n well explored on regression tasks with high dimensional inputs such as images. This paper makes two main contributions that help understand this barely explor ed area. First, we design two new types of cross-category level vision regressio n tasks, namely object discovery and pose estimation of unprecedented complexity in the meta-learning domain for computer vision. To this end, we (i) exhaustive ly evaluate common meta-learning techniques on these tasks, and (ii) quantitativ ely analyze the effect of various deep learning techniques commonly used in rece nt meta-learning algorithms in order to strengthen the generalization capability : data augmentation, domain randomization, task augmentation and meta-regulariza tion. Finally, we (iii) provide some insights and practical recommendations for

training meta-learning algorithms on vision regression tasks. Second, we propose the addition of functional contrastive learning (FCL) over the task representations in Conditional Neural Processes (CNPs) and train in an end-to-end fashion. The experimental results show that the results of prior work are misleading as a consequence of a poor choice of the loss function as well as too small meta-training sets. Specifically, we find that CNPs outperform MAML on most tasks without fine-tuning. Furthermore, we observe that naive task augmentation without a tailored design results in underfitting.

****

Knowledge-Driven Self-Supervised Representation Learning for Facial Action Unit Recognition

Yanan Chang, Shangfei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20417-20426

Facial action unit (AU) recognition is formulated as a supervised learning problem by recent works. However, the complex labeling process makes it challenging to provide AU annotations for large amounts of facial images. To remedy this, we utilize AU labeling rules defined by the Facial Action Coding System (FACS) to design a novel knowledge-driven self-supervised representation learning framework for AU recognition. The representation encoder is trained using large amounts of facial images without AU annotations. AU labeling rules are summarized from FACS to design facial partition manners and determine correlations between facial regions. The method utilizes a backbone network to extract local facial area representations and a project head to map the representations into a low-dimensional latent space. In the latent space, a contrastive learning component leverages the inter-area difference to learn AU-related local representations while maintaining intra-area instance discrimination. Correlations between facial regions summarized from AU labeling rules are also explored to further learn representations using a predicting learning component. Evaluation on two benchmark databases demonstrates that the learned representation is powerful and data-efficient for AU recognition.

****

Selective-Supervised Contrastive Learning With Noisy Labels

Shikun Li, Xiaobo Xia, Shiming Ge, Tongliang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 316-325

Deep networks have strong capacities of embedding data into latent representations and finishing following tasks. However, the capacities largely come from high-quality annotated labels, which are expensive to collect. Noisy labels are more affordable, but result in corrupted representations, leading to poor generalization performance. To learn robust representations and handle noisy labels, we propose selective-supervised contrastive learning (Sel-CL) in this paper. Specifically, Sel-CL extend supervised contrastive learning (Sup-CL), which is powerful in representation learning, but is degraded when there are noisy labels. Sel-CL tackles the direct cause of the problem of Sup-CL. That is, as Sup-CL works in a pair-wise manner, noisy pairs built by noisy labels mislead representation learning. To alleviate the issue, we select confident pairs out of noisy ones for Sup-CL without knowing noise rates. In the selection process, by measuring the agreement between learned representations and given labels, we first identify confident examples that are exploited to build confident pairs. Then, the representation similarity distribution in the built confident pairs is exploited to identify more confident pairs out of noisy pairs. All obtained confident pairs are finally used for Sup-CL to enhance representations. Experiments on multiple noisy datasets demonstrate the robustness of the learned representations by our method, following the state-of-the-art performance. Source codes are available at https://github.com/ShikunLi/Sel-CL.

****

Learning Second Order Local Anomaly for General Face Forgery Detection

Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, Jian Weng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20270-20280

In this work, we propose a novel method to improve the generalization ability of

CNN-based face forgery detectors. Our method considers the feature anomalies of forged faces caused by the prevalent blending operations in face forgery algorithms. Specifically, we propose a weakly supervised Second Order Local Anomaly (SOLA) learning module to mine anomalies in local regions using deep feature maps. SOLA first decomposes the neighborhood of local features by different directions and distances and then calculates the first and second order local anomaly maps which provide more general forgery traces for the classifier. We also propose a Local Enhancement Module (LEM) to improve the discrimination between local features of real and forged regions, so as to ensure accuracy in calculating anomalies. Besides, an improved Adaptive Spatial Rich Model (ASRM) is introduced to help mine subtle noise features via learnable high pass filters. With neither pixel level annotations nor external synthetic data, our method using a simple ResNet18 backbone achieves competitive performances compared with state-of-the-art works when evaluated on unseen forgeries.

****

ADAS: A Direct Adaptation Strategy for Multi-Target Domain Adaptive Semantic Segmentation

Seunghun Lee, Wonhyeok Choi, Changjae Kim, Minwoo Choi, Sunghoon Im; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19196-19206

In this paper, we present a direct adaptation strategy (ADAS), which aims to directly adapt a single model to multiple target domains in a semantic segmentation task without pretrained domain-specific models. To do so, we design a multi-target domain transfer network (MTDT-Net) that aligns visual attributes across domains by transferring the domain distinctive features through a new target adaptive denormalization (TAD) module. Moreover, we propose a bi-directional adaptive region selection (BARS) that reduces the attribute ambiguity among the class labels by adaptively selecting the regions with consistent feature statistics. We show that our single MTDT-Net can synthesize visually pleasing domain transferred images with complex driving datasets, and BARS effectively filters out the unnecessary region of training images for each target domain. With the collaboration of MTDT-Net and BARS, our ADAS achieves state-of-the-art performance for multi-target domain adaptation (MTDA). To the best of our knowledge, our method is the first MTDA method that directly adapts to multiple domains in semantic segmentation.

****

The Devil Is in the Labels: Noisy Label Correction for Robust Scene Graph Generation

Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, Jun Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18869-18878

Unbiased SGG has achieved significant progress over recent years. However, almost all existing SGG models have overlooked the ground-truth annotation qualities of prevailing SGG datasets, i.e., they always assume: 1) all the manually annotated positive samples are equally correct; 2) all the un-annotated negative samples are absolutely background. In this paper, we argue that both assumptions are inapplicable to SGG: there are numerous "noisy" ground-truth predicate labels that break these two assumptions, and these noisy samples actually harm the training of unbiased SGG models. To this end, we propose a novel model-agnostic NoIsy label CorrEction strategy for SGG: NICE. NICE can not only detect noisy samples but also reassign more high-quality predicate labels to them. After the NICE training, we can obtain a cleaner version of SGG dataset for model training. Specifically, NICE consists of three components: negative Noisy Sample Detection (Neg-NSD), positive NSD (Pos-NSD), and Noisy Sample Correction (NSC). Firstly, in Neg-NSD, we formulate this task as an out-of-distribution detection problem, and assign pseudo labels to all detected noisy negative samples. Then, in Pos-NSD, we use a clustering-based algorithm to divide all positive samples into multiple sets, and treat the samples in the noisiest set as noisy positive samples. Lastly, in NSC, we use a simple but effective weighted KNN to reassign new predicate labels to noisy positive samples. Extensive results on different backbones and tas

ks have attested to the effectiveness and generalization abilities of each compo
nent of NICE.
****

LAVT: Language-Aware Vision Transformer for Referring Image Segmentation
Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, Philip H.S. Torr
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 18155-18165
Referring image segmentation is a fundamental vision-language task that aims to
segment out an object referred to by a natural language expression from an image
. One of the key challenges behind this task is leveraging the referring express
ion for highlighting relevant positions in the image. A paradigm for tackling th
is problem is to leverage a powerful vision-language ("cross-modal") decoder to
fuse features independently extracted from a vision encoder and a language encod
er. Recent methods have made remarkable advancements in this paradigm by exploit
ing Transformers as cross-modal decoders, concurrent to the Transformer's overwh
elming success in many other vision-language tasks. Adopting a different approac
h in this work, we show that significantly better cross-modal alignments can be
achieved through the early fusion of linguistic and visual features in intermedi
ate layers of a vision Transformer encoder network. By conducting cross-modal fe
ature fusion in the visual feature encoding stage, we can leverage the well-prov
en correlation modeling power of a Transformer encoder for excavating helpful mu
lti-modal context. This way, accurate segmentation results are readily harvested
 with a light-weight mask predictor. Without bells and whistles, our method surp
asses the previous state-of-the-art methods on RefCOCO, RefCOCO+, and G-Ref by l
arge margins.
****

SimT: Handling Open-Set Noise for Domain Adaptive Semantic Segmentation
Xiaoqing Guo, Jie Liu, Tongliang Liu, Yixuan Yuan; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7032-7041
This paper studies a practical domain adaptative (DA) semantic segmentation prob
lem where only pseudo-labeled target data is accessible through a black-box mode
l. Due to the domain gap and label shift between two domains, pseudo-labeled tar
get data contains mixed closed-set and open-set label noises. In this paper, we
propose a simplex noise transition matrix (SimT) to model the mixed noise distri
butions in DA semantic segmentation and formulate the problem as estimation of S
imT. By exploiting computational geometry analysis and properties of segmentatio
n, we design three complementary regularizers, i.e. volume regularization, ancho
r guidance, convex guarantee, to approximate the true SimT. Specifically, volume
 regularization minimizes the volume of simplex formed by rows of the non-square
 SimT, which ensures outputs of segmentation model to fit into the ground truth
label distribution. To compensate for the lack of open-set knowledge, anchor gui
dance and convex guarantee are devised to facilitate the modeling of open-set no
ise distribution and enhance the discriminative feature learning among closed-se
t and open-set classes. The estimated SimT is further utilized to correct noise
issues in pseudo labels and promote the generalization ability of segmentation m
odel on target domain data. Extensive experimental results demonstrate that the
proposed SimT can be flexibly plugged into existing DA methods to boost the perf
ormance. The source code is available at https://github.com/CityU-AIM-Group/SimT
.
****

Interspace Pruning: Using Adaptive Filter Representations To Improve Training of
 Sparse CNNs
Paul Wimmer, Jens Mehnert, Alexandru Condurache; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12527-12537
Unstructured pruning is well suited to reduce the memory footprint of convolutio
nal neural networks (CNNs), both at training and inference time. CNNs contain pa
rameters arranged in K x K filters. Standard unstructured pruning (SP) reduces t
he memory footprint of CNNs by setting filter elements to zero, thereby specifyi
ng a fixed subspace that constrains the filter. Especially if pruning is applied
 before or during training, this induces a strong bias. To overcome this, we int

roduce interspace pruning (IP), a general tool to improve existing pruning metho ds. It uses filters represented in a dynamic interspace by linear combinations o f an underlying adaptive filter basis (FB). For IP, FB coefficients are set to z ero while un-pruned coefficients and FBs are trained jointly. In this work, we p rovide mathematical evidence for IP's superior performance and demonstrate that IP outperforms SP on all tested state-of-the-art unstructured pruning methods. E specially in challenging situations, like pruning for ImageNet or pruning to hig h sparsity, IP greatly exceeds SP with equal runtime and parameter costs. Finall y, we show that advances of IP are due to improved trainability and superior gen eralization ability.
****

PLAD: Learning To Infer Shape Programs With Pseudo-Labels and Approximate Distri butions
R. Kenny Jones, Homer Walke, Daniel Ritchie; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9871-9880
Inferring programs which generate 2D and 3D shapes is important for reverse engi neering, editing, and more. Training models to perform this task is complicated because paired (shape, program) data is not readily available for many domains, making exact supervised learning infeasible. However, it is possible to get pair ed data by compromising the accuracy of either the assigned program labels or th e shape distribution. Wake-sleep methods use samples from a generative model of shape programs to approximate the distribution of real shapes. In self-training, shapes are passed through a recognition model, which predicts programs that are treated as 'pseudo-labels' for those shapes. Related to these approaches, we in troduce a novel self-training variant unique to program inference, where program pseudo-labels are paired with their executed output shapes, avoiding label mism atch at the cost of an approximate shape distribution. We propose to group these regimes under a single conceptual framework, where training is performed with m aximum likelihood updates sourced from either Pseudo-Labels or an Approximate Di stribution (PLAD). We evaluate these techniques on multiple 2D and 3D shape prog ram inference domains. Compared with policy gradient reinforcement learning, we show that PLAD techniques infer more accurate shape programs and converge signif icantly faster. Finally, we propose to combine updates from different PLAD metho ds within the training of a single model, and find that this approach outperform s any individual technique.
****

PTTR: Relational 3D Point Cloud Object Tracking With Transformer
Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Ha iyu Zhao, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8531-8540
In a point cloud sequence, 3D object tracking aims to predict the location and o rientation of an object in the current search point cloud given a template point cloud. Motivated by the success of transformers, we propose Point Tracking TRan sformer (PTTR), which efficiently predicts high-quality 3D tracking results in a coarse-to-fine manner with the help of transformer operations. PTTR consists of three novel designs. 1) Instead of random sampling, we design Relation-Aware Sa mpling to preserve relevant points to given templates during subsampling. 2) Fur thermore, we propose a Point Relation Transformer (PRT) consisting of a self-att ention and a cross-attention module. The global self-attention operation capture s long-range dependencies to enhance encoded point features for the search area and the template, respectively. Subsequently, we generate the coarse tracking re sults by matching the two sets of point features via cross-attention. 3) Based o n the coarse tracking results, we employ a novel Prediction Refinement Module to obtain the final refined prediction. In addition, we create a large-scale point cloud single object tracking benchmark based on the Waymo Open Dataset. Extensi ve experiments show that PTTR achieves superior point cloud tracking in both acc uracy and efficiency. Our code and dataset will be released upon acceptance.
****

Frequency-Driven Imperceptible Adversarial Attack on Semantic Similarity
Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, Linlin Shen; Proce

edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15315-15324
Current adversarial attack research reveals the vulnerability of learning-based classifiers against carefully crafted perturbations. However, most existing attack methods have inherent limitations in cross-dataset generalization as they rely on a classification layer with a closed set of categories. Furthermore, the perturbations generated by these methods may appear in regions easily perceptible to the human visual system (HVS). To circumvent the former problem, we propose a novel algorithm that attacks semantic similarity on feature representations. In this way, we are able to fool classifiers without limiting attacks to a specific dataset. For imperceptibility, we introduce the low-frequency constraint to limit perturbations within high-frequency components, ensuring perceptual similarity between adversarial examples and originals. Extensive experiments on three datasets (CIFAR-10, CIFAR-100, and ImageNet-1K) and three public online platforms indicate that our attack can yield misleading and transferable adversarial examples across architectures and datasets. Additionally, visualization results and quantitative performance (in terms of four different metrics) show that the proposed algorithm generates more imperceptible perturbations than the state-of-the-art methods. Code is made available at https://github.com/LinQinLiang/SSAH-adversarial-attack.
****

ZZ-Net: A Universal Rotation Equivariant Architecture for 2D Point Clouds
Georg Bökman, Fredrik Kahl, Axel Flinth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10976-10985
In this paper, we are concerned with rotation equivariance on 2D point cloud data. We describe a particular set of functions able to approximate any continuous rotation equivariant and permutation invariant function. Based on this result, we propose a novel neural network architecture for processing 2D point clouds and we prove its universality for approximating functions exhibiting these symmetries. We also show how to extend the architecture to accept a set of 2D-2D correspondences as indata, while maintaining similar equivariance properties. Experiments are presented on the estimation of essential matrices in stereo vision.
****

Video Demoireing With Relation-Based Temporal Consistency
Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17622-17631
Moire patterns, appearing as color distortions, severely degrade the image and video qualities when filming a screen with digital cameras. Considering the increasing demands for capturing videos, we study how to remove such undesirable moire patterns in videos, namely video demoireing. To this end, we introduce the first hand-held video demoireing dataset with a dedicated data collection pipeline to ensure spatial and temporal alignments of captured data. Further, a baseline video demoireing model with implicit feature space alignment and selective feature aggregation is developed to leverage complementary information from nearby frames to improve frame-level video demoireing. More importantly, we propose a relation-based temporal consistency loss to encourage the model to learn temporal consistency priors directly from ground-truth reference videos, which facilitates producing temporally consistent predictions and effectively maintains frame-level qualities. Extensive experiments manifest the superiority of our model. Code is available at https://daipengwa.github.io/VDmoire_ProjectPage/.
****

Co-Domain Symmetry for Complex-Valued Deep Learning
Utkarsh Singhal, Yifei Xing, Stella X. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 681-690
We study complex-valued scaling as a type of symmetry natural and unique to complex-valued measurements and representations. Deep Complex Networks (DCN) extends real-valued algebra to the complex domain without addressing complex-valued scaling. SurReal extends manifold learning to the complex plane, achieving scaling invariance using distances that discard phase information. Treating complex-valu

ed scaling as a co-domain transformation, we design novel equivariant/invariant neural network layer functions and construct architectures that exploit co-domain symmetry. We also propose novel complex-valued representations of RGB images, where complex-valued scaling indicates hue shift or correlated changes across color channels. Benchmarked on MSTAR, CIFAR10, CIFAR100, and SVHN, our co-domain symmetric (CDS) classifiers deliver higher accuracy, better generalization, more robustness to co-domain transformations, and lower model bias and variance than DCN and SurReal with far fewer parameters.
****
Industrial Style Transfer With Large-Scale Geometric Warping and Content Preservation
Jinchao Yang, Fei Guo, Shuo Chen, Jun Li, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7834-7843
We propose a novel style transfer method to quickly create a new visual product with a nice appearance for industrial designers' reference. Given a source product, a target product, and an art style image, our method produces a neural warping field that warps the source shape to imitate the geometric style of the target and a neural texture transformation network that transfers the artistic style to the warped source product. Our model, Industrial Style Transfer (InST), consists of large-scale geometric warping (LGW) and interest-consistency texture transfer (ICTT). LGW aims to explore an unsupervised transformation between the shape masks of the source and target products for fitting large-scale shape warping. Furthermore, we introduce a mask smoothness regularization term to prevent the abrupt changes of the details of the source product. ICTT introduces an interest regularization term to maintain important contents of the warped product when it is stylized by using the art style image. Extensive experimental results demonstrate that InST achieves state-of-the-art performance on multiple visual product design tasks, e.g., companies' snail logos and classical bottles (please see Fig. 1). To the best of our knowledge, we are the first to extend the neural style transfer method to create industrial product appearances. Code is available at https://jcyang98.github.io/InST/home.html
****
Modeling Image Composition for Complex Scene Generation
Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7764-7773
We present a method that achieves state-of-the-art results on challenging (few-shot) layout-to-image generation tasks by accurately modeling textures, structures and relationships contained in a complex scene. After compressing RGB images into patch tokens, we propose the Transformer with Focal Attention (TwFA) for exploring dependencies of object-to-object, object-to-patch and patch-to-patch. Compared to existing CNN-based and Transformer-based generation models that entangled modeling on pixel-level & patch-level and object-level & patch-level respectively, the proposed focal attention predicts the current patch token by only focusing on its highly-related tokens that specified by the spatial layout, thereby achieving disambiguation during training. Furthermore, the proposed TwFA largely increases the data efficiency during training, therefore we propose the first few-shot complex scene generation strategy based on the well-trained TwFA. Comprehensive experiments show the superiority of our method, which significantly increases both quantitative metrics and qualitative visual realism with respect to state-of-the-art CNN-based and transformer-based methods. Code is available at https://github.com/JohnDreamer/TwFA.
****
SS3D: Sparsely-Supervised 3D Object Detection From Point Cloud
Chuandong Liu, Chenqiang Gao, Fangcen Liu, Jiang Liu, Deyu Meng, Xinbo Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8428-8437
Conventional deep learning based methods for 3D object detection require a large amount of 3D bounding box annotations for training, which is expensive to obtai

n in practice. Sparsely annotated object detection, which can largely reduce the annotations, is very challenging since the missingannotated instances would be regarded as the background during training. In this paper, we propose a sparsely supervised 3D object detection method, named SS3D. Aiming to eliminate the negative supervision caused by the missing annotations, we design a missing-annotated instance mining module with strict filtering strategies to mine positive instances. In the meantime, we design a reliable background mining module and a point cloud filling data augmentation strategy to generate the confident data for iteratively learning with reliable supervision. The proposed SS3D is a general framework that can be used to learn any modern 3D object detector. Extensive experiments on the KITTI dataset reveal that on different 3D detectors, the proposed SS3D framework with only 20% annotations required can achieve on-par performance comparing to fullysupervised methods. Comparing with the state-of-the-art semi-supervised 3D objection detection on KITTI, our SS3D improves the benchmarks by significant margins under the same annotation workload. Moreover, our SS3D also outperforms the state-of-the-art weakly-supervised method by remarkable margins, highlighting its effectiveness.

****

Remember the Difference: Cross-Domain Few-Shot Semantic Segmentation via Meta-Memory Transfer

Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7065-7074

Few-shot semantic segmentation intends to predict pixel level categories using only a few labeled samples. Existing few-shot methods focus primarily on the categories sampled from the same distribution. Nevertheless, this assumption cannot always be ensured. The actual domain shift problem significantly reduces the performance of few-shot learning. To remedy this problem, we propose an interesting and challenging cross-domain few-shot semantic segmentation task, where the training and test tasks perform on different domains. Specifically, we first propose a meta-memory bank to improve the generalization of the segmentation network by bridging the domain gap between source and target domains. The meta-memory stores the intra-domain style information from source domain instances and transfers it to target samples. Subsequently, we adopt a new contrastive learning strategy to explore the knowledge of different categories during the training stage. The negative and positive pairs are obtained from the proposed memory-based style augmentation. Comprehensive experiments demonstrate that our proposed method achieves promising results on cross-domain few-shot semantic segmentation tasks on COCO-20, PASCAL-5, FSS-1000, and SUIM datasets.

****

GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation

Yu Deng, Jiaolong Yang, Jianfeng Xiang, Xin Tong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10673-10683

3D-aware image generative modeling aims to generate 3D-consistent images with explicitly controllable camera poses. Recent works have shown promising results by training neural radiance field (NeRF) generators on unstructured 2D images, but still cannot generate highly-realistic images with fine details. A critical reason is that the high memory and computation cost of volumetric representation learning greatly restricts the number of point samples for radiance integration during training. Deficient sampling not only limits the expressive power of the generator to handle fine details but also impedes effective GAN training due to the noise caused by unstable Monte Carlo sampling. We propose a novel approach that regulates point sampling and radiance field learning on 2D manifolds, embodied as a set of learned implicit surfaces in the 3D volume. For each viewing ray, we calculate ray-surface intersections and accumulate their radiance generated by the network. By training and rendering such radiance manifolds, our generator can produce high quality images with realistic fine details and strong visual 3D consistency.

****

UniVIP: A Unified Framework for Self-Supervised Visual Pre-Training
Zhaowen Li, Yousong Zhu, Fan Yang, Wei Li, Chaoyang Zhao, Yingying Chen, Zhiyang Chen, Jiahao Xie, Liwei Wu, Rui Zhao, Ming Tang, Jinqiao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14627-14636
Self-supervised learning (SSL) holds promise in leveraging large amounts of unlabeled data. However, the success of popular SSL methods has limited on single-centric-object images like those in ImageNet and ignores the correlation among the scene and instances, as well as the semantic difference of instances in the scene. To address the above problems, we propose a Unified Self-supervised Visual Pre-training (UniVIP), a novel self-supervised framework to learn versatile visual representations on either single-centric-object or non-iconic dataset. The framework takes into account the representation learning at three levels: 1) the similarity of scene-scene, 2) the correlation of scene-instance, 3) the discrimination of instance-instance. During the learning, we adopt the optimal transport algorithm to automatically measure the discrimination of instances. Massive experiments show that UniVIP pre-trained on non-iconic COCO achieves state-of-the-art transfer performance on a variety of downstream tasks, such as image classification, semi-supervised learning, object detection and segmentation. Furthermore, our method can also exploit single-centric-object dataset such as ImageNet and outperforms BYOL by 2.5% with the same pre-training epochs in linear probing, and surpass current self-supervised object detection methods on COCO dataset, demonstrating its universality and potential.
****
GraFormer: Graph-Oriented Transformer for 3D Pose Estimation
Weixi Zhao, Weiqiang Wang, Yunjie Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20438-20447
In 2D-to-3D pose estimation, it is important to exploit the spatial constraints of 2D joints, but it is not yet well modeled. To better model the relation of joints for 3D pose estimation, we propose an effective but simple network, called GraFormer, where a novel transformer architecture is designed via embedding graph convolution layers after multi-head attention block. The proposed GraFormer is built by repeatedly stacking the GraAttention block and the ChebGConv block. The proposed GraAttention block is a new transformer block designed for processing graph-structured data, which is able to learn better features through capturing global information from all the nodes as well as the explicit adjacency structure of nodes. To model the implicit high-order connection relations among non-neighboring nodes, the ChebGConv block is introduced to exchange information between non-neighboring nodes and attain a larger receptive field. We have empirically shown the superiority of GraFormer through extensive experiments on popular public datasets. Specifically, GraFormer outperforms the state-of-the-art GraghSH on the Human3.6M dataset yet only contains 18% parameters of it
****
Decoupling Zero-Shot Semantic Segmentation
Jian Ding, Nan Xue, Gui-Song Xia, Dengxin Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11583-11592
Zero-shot semantic segmentation (ZS3) aims to segment the novel categories that have not been seen in the training. Existing works formulate ZS3 as a pixel-level zero-shot classification problem, and transfer semantic knowledge from seen classes to unseen ones with the help of language models pre-trained only with texts. While simple, the pixel-level ZS3 formulation shows the limited capability to integrate vision-language models that are often pre-trained with image-text pairs and currently demonstrate great potential for vision tasks. Inspired by the observation that humans often perform segment-level semantic labeling, we propose to decouple the ZS3 into two sub-tasks: 1) a class-agnostic grouping task to group the pixels into segments. 2) a zero-shot classification task on segments. The former task does not involve category information and can be directly transferred to group pixels for unseen classes. The latter task performs at segment-level and provides a natural way to leverage large-scale vision-language models pre-trained with image-text pairs (e.g. CLIP) for ZS3. Based on the decoupling formu

lation, we propose a simple and effective zero-shot semantic segmentation model, called ZegFormer, which outperforms the previous methods on ZS3 standard benchmarks by large margins, e.g., 22 points on the PAS-CAL VOC and 3 points on the COCO-Stuff in terms of mIoU for unseen classes. Code will be released at https://github.com/dingjiansw101/ZegFormer.

****

Neural Collaborative Graph Machines for Table Structure Recognition

Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, Bo Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4533-4542

Recently, table structure recognition has achieved impressive progress with the help of deep graph models. Most of them exploit single visual cues of tabular elements or simply combine visual cues with other modalities via early fusion to reason their graph relationships. However, neither early fusion nor individually reasoning in terms of multiple modalities can be appropriate for all varieties of table structures with great diversity. Instead, different modalities are expected to collaborate with each other in different patterns for different table cases. In the community, the importance of intra-inter modality interactions for table structure reasoning is still unexplored. In this paper, we define it as heterogeneous table structure recognition (Hetero-TSR) problem. With the aim of filling this gap, we present a novel Neural Collaborative Graph Machines (NCGM) equipped with stacked collaborative blocks, which alternatively extracts intra-modality context and models inter-modality interactions in a hierarchical way. It can represent the intra-inter modality relationships of tabular elements more robustly, which significantly improves the recognition performance. We also show that the proposed NCGM can modulate collaborative pattern of different modalities conditioned on the context of intra-modality cues, which is vital for diversified table cases. Experimental results on benchmarks demonstrate our proposed NCGM achieves state-of-the-art performance and beats other contemporary methods by a large margin especially under challenging scenarios.

****

Towards Robust Vision Transformer

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, Hui Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12042-12051

Recent advances on Vision Transformer (ViT) and its improved variants have shown that self-attention-based networks surpass traditional Convolutional Neural Networks (CNNs) in most vision tasks. However, existing ViTs focus on the standard accuracy and computation cost, lacking the investigation of the intrinsic influence on model robustness and generalization. In this work, we conduct systematic evaluation on components of ViTs in terms of their impact on robustness to adversarial examples, common corruptions and distribution shifts. We find some components can be harmful to robustness. By leveraging robust components as building blocks of ViTs, we propose Robust Vision Transformer (RVT), which is a new vision transformer and has superior performance with strong robustness. Inspired by the findings during the evaluation, we further propose two new plug-and-play techniques called position-aware attention scaling and patch-wise augmentation to augment our RVT, which we abbreviate as RVT*. The experimental results of RVT on ImageNet and six robustness benchmarks demonstrate its advanced robustness and generalization ability compared with previous ViTs and state-of-the-art CNNs. Furthermore, RVT-S* achieves Top-1 rank on multiple robustness leaderboards including ImageNet-C, ImageNet-Sketch and ImageNet-R.

****

DeepCurrents: Learning Implicit Representations of Shapes With Boundaries

David Palmer, Dmitriy Smirnov, Stephanie Wang, Albert Chern, Justin Solomon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18665-18675

Recent techniques have been successful in reconstructing surfaces as level sets of learned functions (such as signed distance fields) parameterized by deep neural networks. Many of these methods, however, learn only closed surfaces and are

unable to reconstruct shapes with boundary curves. We propose a hybrid shape rep
resentation that combines explicit boundary curves with implicit learned interio
rs. Using machinery from geometric measure theory, we parameterize currents usin
g deep networks and use stochastic gradient descent to solve a minimal surface p
roblem. By modifying the metric according to target geometry coming, e.g., from
a mesh or point cloud, we can use this approach to represent arbitrary surfaces,
 learning implicitly defined shapes with explicitly defined boundary curves. We
further demonstrate learning families of shapes jointly parameterized by boundar
y curves and latent codes.
****

Learning Affordance Grounding From Exocentric Images
Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, Dacheng Tao; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2
252-2261
Affordance grounding, a task to ground (i.e., localize) action possibility regio
n in objects, which faces the challenge of establishing an explicit link with ob
ject parts due to the diversity of interactive affordance. Human has the ability
 that transform the various exocentric interactions to invariant egocentric affo
rdance so as to counter the impact of interactive diversity. To empower an agent
 with such ability, this paper proposes a task of affordance grounding from exoc
entric view, i.e., given exocentric human-object interaction and egocentric obje
ct images, learning the affordance knowledge of the object and transferring it t
o the egocentric image using only the affordance label as supervision. To this e
nd, we devise a cross-view knowledge transfer framework that extracts affordance
-specific features from exocentric interactions and enhances the perception of a
ffordance regions by preserving affordance correlation. Specifically, an Afforda
nce Invariance Mining module is devised to extract specific clues by minimizing
the intra-class differences originated from interaction habits in exocentric ima
ges. Besides, an Affordance Co-relation Preserving strategy is presented to perc
eive and localize affordance by aligning the co-relation matrix of predicted res
ults between the two views. Particularly, an affordance grounding dataset named
AGD20K is constructed by collecting and labeling over 20K images from 36 afforda
nce categories. Experimental results demonstrate that our method outperforms the
 representative models in terms of objective metrics and visual quality. Code: g
ithub.com/lhc1224/Cross-View-AG.
****

Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects
 and Robustness to Occlusions
Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, Vincent Lepetit; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 6771-6780
We present a method that can recognize new objects and estimate their 3D pose in
 RGB images even under partial occlusions. Our method requires neither a trainin
g phase on these objects nor real images depicting them, only their CAD models.
It relies on a small set of training objects to learn local object representatio
ns, which allow us to locally match the input image to a set of "templates", ren
dered images of the CAD models for the new objects. In contrast with the state-o
f-the-art methods, the new objects on which our method is applied can be very di
fferent from the training objects. As a result, we are the first to show general
ization without retraining on the LINEMOD and Occlusion-LINEMOD datasets. Our an
alysis of the failure modes of previous template-based approaches further confir
ms the benefits of local features for template matching. We outperform the state
-of-the-art template matching methods on the LINEMOD, Occlusion-LINEMOD and T-LE
SS datasets. Our source code and data are publicly available at https://github.c
om/nv-nguyen/template-pose
****

Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversa
rial Transferability
Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, Kun He; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,

pp. 14983-14992
The black-box adversarial attack has attracted impressive attention for its practical use in the field of deep learning security. Meanwhile, it is very challenging as there is no access to the network architecture or internal weights of the target model. Based on the hypothesis that if an example remains adversarial for multiple models, then it is more likely to transfer the attack capability to other models, the ensemble-based adversarial attack methods are efficient and widely used for black-box attacks. However, ways of ensemble attack are rather less investigated, and existing ensemble attacks simply fuse the outputs of all the models evenly. In this work, we treat the iterative ensemble attack as a stochastic gradient descent optimization process, in which the variance of the gradients on different models may lead to poor local optima. To this end, we propose a novel attack method called the stochastic variance reduced ensemble (SVRE) attack, which could reduce the gradient variance of the ensemble models and take full advantage of the ensemble attack. Empirical results on the standard ImageNet dataset demonstrate that the proposed method could boost the adversarial transferability and outperforms existing ensemble attacks significantly. Code is available at https://github.com/JHL-HUST/SVRE.
****

Unknown-Aware Object Detection: Learning What You Don't Know From Videos in the Wild
Xuefeng Du, Xin Wang, Gabriel Gozum, Yixuan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13678-13688
Building reliable object detectors that can detect out-of-distribution (OOD) objects is critical yet underexplored. One of the key challenges is that models lack supervision signals from unknown data, producing overconfident predictions on OOD objects. We propose a new unknown-aware object detection framework through Spatial-Temporal Unknown Distillation (STUD), which distills unknown objects from videos in the wild and meaningfully regularizes the model's decision boundary. STUD first identifies the unknown candidate object proposals in the spatial dimension, and then aggregates the candidates across multiple video frames to form a diverse set of unknown objects near the decision boundary. Alongside, we employ an energy-based uncertainty regularization loss, which contrastively shapes the uncertainty space between the in-distribution and distilled unknown objects. STUD establishes the state-of-the-art performance on OOD detection tasks for object detection, reducing the FPR95 score by over 10% compared to the previous best method.
****

Multi-Modal Extreme Classification
Anshul Mittal, Kunal Dahiya, Shreya Malani, Janani Ramaswamy, Seba Kuruvilla, Jitendra Ajmera, Keng-hao Chang, Sumeet Agarwal, Purushottam Kar, Manik Varma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12393-12402
This paper develops the MUFIN technique for extreme classification (XC) tasks with millions of labels where datapoints and labels are endowed with visual and textual descriptors. Applications of MUFIN to product-to-product recommendation and bid query prediction over several millions of products are presented. Contemporary multi-modal methods frequently rely on purely embedding-based methods. On the other hand, XC methods utilize classifier architectures to offer superior accuracies than embedding-only methods but mostly focus on text-based categorization tasks. MUFIN bridges this gap by reformulating multi-modal categorization as an XC problem with several millions of labels. This presents the twin challenges of developing multi-modal architectures that can offer embeddings sufficiently expressive to allow accurate categorization over millions of labels; and training and inference routines that scale logarithmically in the number of labels. MUFIN develops an architecture based on cross-modal attention and trains it in a modular fashion using pre-training and positive and negative mining. A novel product-to-product recommendation dataset MM-AmazonTitles-300K containing over 300K products was curated from publicly available amazon.com listings with each product endowed with a title and multiple images. On the MM-AmazonTitles-300K and Polyv

ore datasets, and a dataset with over 4 million labels curated from click logs o f the Bing search engine, MUFIN offered at least 3% higher accuracy than leading text-based, image-based and multi-modal techniques.
****

IFOR: Iterative Flow Minimization for Robotic Object Rearrangement
Ankit Goyal, Arsalan Mousavian, Chris Paxton, Yu-Wei Chao, Brian Okorn, Jia Deng , Dieter Fox; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2022, pp. 14787-14797
Accurate object rearrangement from vision is a crucial problem for a wide variet y of real-world robotics applications in unstructured environments. We propose I FOR, Iterative Flow Minimization for Robotic Object Rearrangement, an end-to-end method for the challenging problem of object rearrangement for unknown objects given an RGBD image of the original and final scenes. First, we learn an optical flow model based on RAFT to estimate the relative transformation of the objects purely from synthetic data. This flow is then used in an iterative minimization algorithm to achieve accurate positioning of previously unseen objects. Crucial ly, we show that our method applies to cluttered scenes, and in the real world, while training only on synthetic data. Videos are available at https://imankgoya l.github.io/ifor.html.
****

Training-Free Transformer Architecture Search
Qinqin Zhou, Kekai Sheng, Xiawu Zheng, Ke Li, Xing Sun, Yonghong Tian, Jie Chen, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2022, pp. 10894-10903
Recently, Vision Transformer (ViT) has achieved remarkable success in several co mputer vision tasks. The progresses are highly relevant to the architecture desi gn, then it is worthwhile to propose Transformer Architecture Search (TAS) to se arch for better ViTs automatically. However, current TAS methods are time-consum ing and existing zero-cost proxies in CNN do not generalize well to the ViT sear ch space according to our experimental observations. In this paper, for the firs t time, we investigate how to conduct TAS in a training-free manner and devise a n effective training-free TAS (TF-TAS) scheme. Firstly, we observe that the prop erties of multi-head self-attention (MSA) and multi-layer perceptron (MLP) in Vi Ts are quite different and that the synaptic diversity of MSA affects the perfor mance notably. Secondly, based on the observation, we devise a modular strategy in TF-TAS that evaluates and ranks ViT architectures from two theoretical perspe ctives: synaptic diversity and synaptic saliency, termed as DSS-indicator. With DSS-indicator, evaluation results are strongly correlated with the test accuraci es of ViT models. Experimental results demonstrate that our TF-TAS achieves a co mpetitive performance against the state-of-the-art manually or automatically des ign ViT architectures, and it promotes the searching efficiency in ViT search sp ace greatly: from about 24 GPU days to less than 0.5 GPU days. Moreover, the pro posed DSS-indicator outperforms the existing cutting-edge zero-cost approaches (e.g., TE-score and NASWOT).
****

Zero Experience Required: Plug & Play Modular Transfer Learning for Semantic Vis ual Navigation
Ziad Al-Halah, Santhosh Kumar Ramakrishnan, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17031-17041
In reinforcement learning for visual navigation, it is common to develop a model for each new task, and train that model from scratch with task-specific interac tions in 3D environments. However, this process is expensive; massive amounts of interactions are needed for the model to generalize well. Moreover, this proces s is repeated whenever there is a change in the task type or the goal modality. We present a unified approach to visual navigation using a novel modular transfe r learning model. Our model can effectively leverage its experience from one sou rce task and apply it to multiple target tasks (e.g., ObjectNav, RoomNav, ViewNa v) with various goal modalities (e.g., image, sketch, audio, label). Furthermore , our model enables zero-shot experience learning, whereby it can solve the targ

et tasks without receiving any task-specific interactive training. Our experimen ts on multiple photorealistic datasets and challenging tasks show that our appro ach learns faster, generalizes better, and outperforms SoTA models by a signific ant margin. Project page: https://vision.cs.utexas.edu/projects/zsel/
****

Non-Isotropy Regularization for Proxy-Based Deep Metric Learning
Karsten Roth, Oriol Vinyals, Zeynep Akata; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7420-7430
Deep Metric Learning (DML) aims to learn representation spaces on which semantic relations can simply be expressed through predefined distance metrics. Best per forming approaches commonly leverage class proxies as sample stand-ins for bette r convergence and generalization. However, these proxy-methods solely optimize f or sample-proxy distances. Given the inherent non-bijectiveness of used distance functions, this can induce locally isotropic sample distributions, leading to c rucial semantic context being missed due to difficulties resolving local structu res and intraclass relations between samples. To alleviate this problem, we prop ose non-isotropy regularization (NIR) for proxy-based Deep Metric Learning. By l everaging Normalizing Flows, we enforce unique translatability of samples from t heir respective class proxies. This allows us to explicitly induce a non-isotrop ic distribution of samples around a proxy to optimize for. In doing so, we equip proxy-based objectives to better learn local structures. Extensive experiments highlight consistent generalization benefits of NIR while achieving competitive and state-of-the-art performance on the standard benchmarks CUB200-2011, Cars196 and Stanford Online Products. In addition, we find the superior convergence pro perties of proxy-based methods to still be retained or even improved, making NIR very attractive for practical usage. Code available at github.com/ExplainableML /NonIsotropicProxyDML.
****

C2AM: Contrastive Learning of Class-Agnostic Activation Map for Weakly Supervise d Object Localization and Semantic Segmentation
Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, Linlin Sh en; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 989-998
While class activation map (CAM) generated by image classification network has b een widely used for weakly supervised object localization (WSOL) and semantic se gmentation (WSSS), such classifiers usually focus on discriminative object regio ns. In this paper, we propose Contrastive learning for Class-agnostic Activation Map (C^2AM) generation only using unlabeled image data, without the involvement of image-level supervision. The core idea comes from the observation that i) se mantic information of foreground objects usually differs from their backgrounds; ii) foreground objects with similar appearance or background with similar color /texture have similar representations in the feature space. We form the positive and negative pairs based on the above relations and force the network to disent angle foreground and background with a class-agnostic activation map using a nov el contrastive loss. As the network is guided to discriminate cross-image foregr ound-background, the class-agnostic activation maps learned by our approach gene rate more complete object regions. We successfully extracted from C^2AM class-ag nostic object bounding boxes for object localization and background cues to refi ne CAM generated by classification network for semantic segmentation. Extensive experiments on CUB-200-2011, ImageNet-1K, and PASCAL VOC2012 datasets show that both WSOL and WSSS can benefit from the proposed C^2AM.
****

TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation
Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12083-12093
Although vision transformers (ViTs) have achieved great success in computer visi on, the heavy computational cost hampers their applications to dense prediction tasks such as semantic segmentation on mobile devices. In this paper, we present a mobile-friendly architecture named Token Pyramid Vision Transformer (TopForme

r). The proposed TopFormer takes Tokens from various scales as input to produce scale-aware semantic features, which are then injected into the corresponding tokens to augment the representation. Experimental results demonstrate that our method significantly outperforms CNN- and ViT-based networks across several semantic segmentation datasets and achieves a good trade-off between accuracy and latency. On the ADE20K dataset, TopFormer achieves 5% higher accuracy in mIoU than MobileNetV3 with lower latency on an ARM-based mobile device. Furthermore, the tiny version of TopFormer achieves real-time inference on an ARM-based mobile device with competitive results. The code and models are available at https://github.com/hustvl/TopFormer.
****

3DAC: Learning Attribute Compression for Point Clouds
Guangchi Fang, Qingyong Hu, Hanyun Wang, Yiling Xu, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14819-14828
We study the problem of attribute compression for large-scale unstructured 3D point clouds. Through an in-depth exploration of the relationships between different encoding steps and different attribute channels, we introduce a deep compression network, termed 3DAC, to explicitly compress the attributes of 3D point clouds and reduce storage usage in this paper. Specifically, the point cloud attributes such as color and reflectance are firstly converted to transform coefficients. We then propose a deep entropy model to model the probabilities of these coefficients by considering information hidden in attribute transforms and previous encoded attributes. Finally, the estimated probabilities are used to further compress these transform coefficients to a final attributes bitstream. Extensive experiments conducted on both indoor and outdoor large-scale open point cloud datasets, including ScanNet and SemanticKITTI, demonstrated the superior compression rates and reconstruction quality of the proposed 3DAC.
****

Learning a Structured Latent Space for Unsupervised Point Cloud Completion
Yingjie Cai, Kwan-Yee Lin, Chao Zhang, Qiang Wang, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5543-5553
Unsupervised point cloud completion aims at estimating the corresponding complete point cloud of a partial point cloud in an unpaired manner. It is a crucial but challenging problem since there is no paired partial-complete supervision that can be exploited directly. In this work, we propose a novel framework, which learns a unified and structured latent space that encoding both partial and complete point clouds. Specifically, we map a series of related partial point clouds into multiple complete shape and occlusion code pairs and fuse the codes to obtain their representations in the unified latent space. To enforce the learning of such a structured latent space, the proposed method adopts a series of constraints including structured ranking regularization, latent code swapping constraint, and distribution supervision on the related partial point clouds. By establishing such a unified and structured latent space, better partial-complete geometry consistency and shape completion accuracy can be achieved. Extensive experiments show that our proposed method consistently outperforms state-of-the-art unsupervised methods on both synthetic ShapeNet and real-world KITTI, ScanNet, and Matterport3D datasets.
****

The Wanderings of Odysseus in 3D Scenes
Yan Zhang, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20481-20491
Our goal is to populate digital environments, in which digital humans have diverse body shapes, move perpetually, and have plausible body-scene contact. The core challenge is to generate realistic, controllable, and infinitely long motions for diverse 3D bodies. To this end, we propose generative motion primitives via body surface markers, or GAMMA in short. In our solution, we decompose the long-term motion into a time sequence of motion primitives. We exploit body surface markers and conditional variational autoencoder to model each motion primitive, a

nd generate long-term motion by implementing the generative model recursively. To control the motion to reach a goal, we apply a policy network to explore the generative model's latent space and use a tree-based search to preserve the motion quality during testing. Experiments show that our method can produce more realistic and controllable motion than state-of-the-art data-driven methods. With conventional path-finding algorithms, the generated human bodies can realistically move long distances for a long period of time in the scene. Code is released for research purposes at: https://yz-cnsdqz.github. io/eigenmotion/GAMMA/
****

Few-Shot Learning With Noisy Labels
Kevin J. Liang, Samrudhdhi B. Rangrej, Vladan Petrovic, Tal Hassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9089-9098
Few-shot learning (FSL) methods typically assume clean support sets with accurately labeled samples when training on novel classes. This assumption can often be unrealistic: support sets, no matter how small, can still include mislabeled samples. Robustness to label noise is therefore essential for FSL methods to be practical, but this problem surprisingly remains largely unexplored. To address mislabeled samples in FSL settings, we make several technical contributions. (1) We offer simple, yet effective, feature aggregation methods, improving the prototypes used by ProtoNet, a popular FSL technique. (2) We describe a novel Transformer model for Noisy Few-Shot Learning (TraNFS). TraNFS leverages a transformer's attention mechanism to weigh mislabeled versus correct samples. (3) Finally, we extensively test these methods on noisy versions of MiniImageNet and TieredImageNet. Our results show that TraNFS is on-par with leading FSL methods on clean support sets, yet outperforms them, by far, in the presence of label noise.
****

Understanding 3D Object Articulation in Internet Videos
Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, David F. Fouhey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1599-1609
We propose to investigate detecting and characterizing the 3D planar articulation of objects from ordinary RGB videos. While seemingly easy for humans, this problem poses many challenges for computers. Our approach is based on a top-down detection system that finds planes that can be articulated. This approach is followed by optimizing for a 3D plane that explains a sequence of detected articulations. We show that this system can be trained on a combination of videos and 3D scan datasets. When tested on a dataset of challenging Internet videos and the Charades dataset, our approach obtains strong performance.
****

Multi-Level Representation Learning With Semantic Alignment for Referring Video Object Segmentation
Dongming Wu, Xingping Dong, Ling Shao, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4996-5005
Referring video object segmentation (RVOS) is a challenging language-guided video grounding task, which requires comprehensively understanding the semantic information of both video content and language queries for object prediction. However, existing methods adopt multi-modal fusion at a frame-based spatial granularity. The limitation of visual representation is prone to causing vision-language mismatching and producing poor segmentation results. To address this, we propose a novel multi-level representation learning approach, which explores the inherent structure of the video content to provide a set of discriminative visual embedding, enabling more effective vision-language semantic alignment. Specifically, we embed different visual cues in terms of visual granularity, including multi-frame long-temporal information at video level, intra-frame spatial semantics at frame level, and enhanced object-aware feature prior at object level. With the powerful multi-level visual embedding and carefully-designed dynamic alignment, our model can generate a robust representation for accurate video object segmentation. Extensive experiments on Refer-DAVIS_ 17 and Refer-YouTube-VOS demonstrat

e that our model achieves superior performance both in segmentation accuracy and inference speed.

****

Paramixer: Parameterizing Mixing Links in Sparse Factors Works Better Than Dot-Product Self-Attention

Tong Yu, Ruslan Khalitov, Lei Cheng, Zhirong Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 691-700

Self-Attention is a widely used building block in neural modeling to mix long-range data elements. Most self-attention neural networks employ pairwise dot-products to specify the attention coefficients. However, these methods require O(N^2) computing cost for sequence length N. Even though some approximation methods have been introduced to relieve the quadratic cost, the performance of the dot-product approach is still bottlenecked by the low-rank constraint in the attention matrix factorization. In this paper, we propose a novel scalable and effective mixing building block called Paramixer. Our method factorizes the interaction matrix into several sparse matrices, where we parameterize the non-zero entries by MLPs with the data elements as input. The overall computing cost of the new building block is as low as O(N \log N). Moreover, all factorizing matrices in Paramixer are full-rank, so it does not suffer from the low-rank bottleneck. We have tested the new method on both synthetic and various real-world long sequential data sets and compared it with several state-of-the-art attention networks. The experimental results show that Paramixer has better performance in most learning tasks.

****

Interactive Image Synthesis With Panoptic Layout Generation

Bo Wang, Tao Wu, Minfeng Zhu, Peng Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7783-7792

Interactive image synthesis from user-guided input is a challenging task when users wish to control the scene structure of a generated image with ease. Although remarkable progress has been made on layout-based image synthesis approaches, existing methods require high-precision inputs such as accurately placed bounding boxes, which might be constantly violated in an interactive setting. When placement of bounding boxes is subject to perturbation, layout-based models suffer from "missing regions" in the constructed semantic layouts and hence undesirable artifacts in the generated images. In this work, we propose Panoptic Layout Generative Adversarial Network (PLGAN) to address this challenge. The PLGAN employs panoptic theory which distinguishes object categories between "stuff" with amorphous boundaries and "things" with well-defined shapes, such that stuff and instance layouts are constructed through separate branches and later fused into panoptic layouts. In particular, the stuff layouts can take amorphous shapes and fill up the missing regions left out by the instance layouts. We experimentally compare our PLGAN with state-of-the-art layout-based models on the COCO-Stuff, Visual Genome, and Landscape datasets. The advantages of PLGAN are not only visually demonstrated but quantitatively verified in terms of inception score, Frechet inception distance, classification accuracy score, and coverage.

****

Pseudo-Stereo for Monocular 3D Object Detection in Autonomous Driving

Yi-Nan Chen, Hang Dai, Yong Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 887-897

Pseudo-LiDAR 3D detectors have made remarkable progress in monocular 3D detection by enhancing the capability of perceiving depth with depth estimation networks, and using LiDAR-based 3D detection architectures. The advanced stereo 3D detectors can also accurately localize 3D objects. The gap in image-to-image generation for stereo views is much smaller than that in image-to-LiDAR generation. Motivated by this, we propose a Pseudo-Stereo 3D detection framework with three novel virtual view generation methods, including image-level generation, feature-level generation, and feature-clone, for detecting 3D objects from a single image. Our analysis of depth-aware learning shows that the depth loss is effective in only feature-level virtual view generation and the estimated depth map is effective in both image-level and feature-level in our framework. We propose a disparit

y-wise dynamic convolution with dynamic kernels sampled from the disparity featu re map to filter the features adaptively from a single image for generating virt ual image features, which eases the feature degradation caused by the depth esti mation errors. Till submission (November 18, 2021), our Pseudo-Stereo 3D detecti on framework ranks 1st on car, pedestrian, and cyclist among the monocular 3D de tectors with publications on the KITTI-3D benchmark. Our code will be released.
****

All-in-One Image Restoration for Unknown Corruption
Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17452-17462
In this paper, we study a challenging problem in image restoration, namely, how to develop an all-in-one method that could recover images from a variety of unkn own corruption types and levels. To this end, we propose an All-in-one Image Res toration Network (AirNet) consisting of two neural modules, named Contrastive-Ba sed Degraded Encoder (CBDE) and Degradation-Guided Restoration Network (DGRN). T he major advantages of AirNet are two-fold. First, it is an all-in-one solution which could recover various degraded images in one network. Second, AirNet is fr ee from the prior of the corruption types and levels, which just uses the observ ed corrupted image to perform inference. These two advantages enable AirNet to e njoy better flexibility and higher economy in real world scenarios wherein the p riors on the corruptions are hard to know and the degradation will change with s pace and time. Extensive experimental results show the proposed method outperfor ms 17 image restoration baselines on four challenging datasets. The code is avai lable at https://github.com/XLearning-SCU/2022-CVPR-AirNet.
****

Syntax-Aware Network for Handwritten Mathematical Expression Recognition
Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2022, pp. 4553-4562
Handwritten mathematical expression recognition (HMER) is a challenging task tha t has many potential applications. Recent methods for HMER have achieved outstan ding performance with an encoder-decoder architecture. However, these methods ad here to the paradigm that the prediction is made "from one character to another" , which inevitably yields prediction errors due to the complicated structures of mathematical expressions or crabbed handwritings. In this paper, we propose a s imple and efficient method for HMER, which is the first to incorporate syntax in formation into an encoder-decoder network. Specifically, we present a set of gra mmar rules for converting the LaTeX markup sequence of each expression into a pa rsing tree; then, we model the markup sequence prediction as a tree traverse pro cess with a deep neural network. In this way, the proposed method can effectivel y describe the syntax context of expressions, alleviating the structure predicti on errors of HMER. Experiments on three benchmark datasets demonstrate that our method achieves better recognition performance than prior arts. To further valid ate the effectiveness of our method, we create a large-scale dataset consisting of 100k handwritten mathematical expression images acquired from ten thousand wr iters. The source code, new dataset, and pre-trained models of this work will be publicly available.
****

Sketching Without Worrying: Noise-Tolerant Sketch-Based Image Retrieval
Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sai n, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 999-1008
Sketching enables many exciting applications, notably, image retrieval. The fear -to-sketch problem (i.e., "I can't sketch") has however proven to be fatal for i ts widespread adoption. This paper tackles this "fear" head on, and for the firs t time, proposes an auxiliary module for existing retrieval models that predomin antly lets the users sketch without having to worry. We first conducted a pilot study that revealed the secret lies in the existence of noisy strokes, but not s o much of the "I can't sketch". We consequently design a stroke subset selector

that detects noisy strokes, leaving only those which make a positive contributio n towards successful retrieval. Our Reinforcement Learning based formulation qua ntifies the importance of each stroke present in a given subset, based on the ex tent to which that stroke contributes to retrieval. When combined with pre-train ed retrieval models as a pre-processing module, we achieve a significant gain of 8%-10% over standard baselines and in turn report new state-of-the-art performa nce. Last but not least, we demonstrate the selector once trained, can also be u sed in a plug-and-play manner to empower various sketch applications in ways tha t were not previously possible.

****

PUMP: Pyramidal and Uniqueness Matching Priors for Unsupervised Learning of Loca l Descriptors

Jérome Revaud, Vincent Leroy, Philippe Weinzaepfel, Boris Chidlovskii; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3926-3936

Existing approaches for learning local image descriptors have shown remarkable a chievements in a wide range of geometric tasks. However, most of them require pe r-pixel correspondence-level supervision, which is difficult to acquire at scale and in high quality. In this paper, we propose to explicitly integrate two matc hing priors in a single loss in order to learn local descriptors without supervi sion. Given two images depicting the same scene, we extract pixel descriptors an d build a correlation volume. The first prior enforces the local consistency of matches in this volume via a pyramidal structure iteratively constructed using a non-parametric module. The second prior exploits the fact that each descriptor should match with at most one descriptor from the other image. We combine our un supervised loss with a standard self-supervised loss trained from synthetic imag e augmentations. Feature descriptors learned by the proposed approach outperform their fully- and self-supervised counterparts on various geometric benchmarks s uch as visual localization and image matching, achieving state-of-the-art perfor mance. Project webpage: https://europe.naverlabs.com/research/3d-vision/pump

****

PlanarRecon: Real-Time 3D Plane Detection and Reconstruction From Posed Monocula r Videos

Yiming Xie, Matheus Gadelha, Fengting Yang, Xiaowei Zhou, Huaizu Jiang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2022, pp. 6219-6228

We present PlanarRecon -- a novel framework for globally coherent detection and reconstruction of 3D planes from a posed monocular video. Unlike previous works that detect planes in 2D from a single image, PlanarRecon incrementally detects planes in 3D for each video fragment, which consists of a set of key frames, fro m a volumetric representation of the scene using neural networks. A learning-bas ed tracking and fusion module is designed to merge planes from previous fragment s to form a coherent global plane reconstruction. Such design allows PlanarRecon to integrate observations from multiple views within each fragment and temporal information across different ones, resulting in an accurate and coherent recons truction of the scene abstraction with low-polygonal geometry. Experiments show that the proposed approach achieves state-of-the-art performances on the ScanNet dataset while being real-time.

****

Deep Equilibrium Optical Flow Estimation

Shaojie Bai, Zhengyang Geng, Yash Savani, J. Zico Kolter; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 62 0-630

Many recent state-of-the-art (SOTA) optical flow models use finite-step recurren t update operations to emulate traditional algorithms by encouraging iterative r efinements toward a stable flow estimation. However, these RNNs impose large com putation and memory overheads, and are not directly trained to model such "stabl e estimation". They can converge poorly and thereby suffer from performance degr adation. To combat these drawbacks, we propose deep equilibrium (DEQ) flow estim ators, an approach that directly solves for the flow as the infinite-level fixed

point of an implicit layer (using any black-box solver), and differentiates thr
ough this fixed point analytically (thus requiring O(1) training memory). This i
mplicit-depth approach is not predicated on any specific model, and thus can be
applied to a wide range of SOTA flow estimation model designs (e.g., RAFT and GM
A). The use of these DEQ flow estimators allows us to compute the flow faster us
ing, e.g., fixed-point reuse and inexact gradients, consumes 4-6x less training
memory than the recurrent counterpart, and achieves better results with the same
computation budget. In addition, we propose a novel, sparse fixed-point correct
ion scheme to stabilize our DEQ flow estimators, which addresses a longstanding
challenge for DEQ models in general. We test our approach in various realistic s
ettings and show that it improves SOTA methods on Sintel and KITTI datasets with
substantially better computational and memory efficiency.
****

Optimizing Video Prediction via Video Frame Interpolation
Yue Wu, Qiang Wen, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2022, pp. 17814-17823
Video prediction is an extrapolation task that predicts future frames given past
frames, and video frame interpolation is an interpolation task that estimates i
ntermediate frames between two frames. We have witnessed the tremendous advancem
ent of video frame interpolation, but the general video prediction in the wild i
s still an open question. Inspired by the photo-realistic results of video frame
interpolation, we present a new optimization framework for video prediction via
video frame interpolation, in which we solve an extrapolation problem based on
an interpolation model. Our video prediction framework is based on optimization
with a pretrained differentiable video frame interpolation module without the ne
ed for a training dataset, and thus there is no domain gap issue between trainin
g and test data. Also, our approach does not need any additional information suc
h as semantic or instance maps, which makes our framework applicable to any vide
o. Extensive experiments on the Cityscapes, KITTI, DAVIS, Middlebury, and Vimeo9
0K datasets show that our video prediction results are robust in general scenari
os, and our approach outperforms other video prediction methods that require a l
arge amount of training data or extra semantic information.
****

Motron: Multimodal Probabilistic Human Motion Forecasting
Tim Salzmann, Marco Pavone, Markus Ryll; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6457-6466
Autonomous systems and humans are increasingly sharing the same space. Robots wo
rk side by side or even hand in hand with humans to balance each other's limitat
ions. Such cooperative interactions are ever more sophisticated. Thus, the abili
ty to reason not just about a human's center of gravity position, but also its g
ranular motion is an important prerequisite for human-robot interaction. Though,
many algorithms ignore the multimodal nature of humans or neglect uncertainty i
n their motion forecasts. We present Motron, a multimodal, probabilistic, graph-
structured model, that captures human's multimodality using probabilistic method
s while being able to output deterministic maximum-likelihood motions and corres
ponding confidence values for each mode. Our model aims to be tightly integrated
with the robotic planning-control-interaction loop; outputting physically feasi
ble human motions and being computationally efficient. We demonstrate the perfor
mance of our model on several challenging real-world motion forecasting datasets
, outperforming a wide array of generative/variational methods while providing s
tate-of-the-art single-output motions if required. Both using significantly less
computational power than state-of-the art algorithms.
****

Episodic Memory Question Answering
Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv
Batra, Devi Parikh; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2022, pp. 19119-19128
Egocentric augmented reality devices such as wearable glasses passively capture
visual data as a human wearer tours a home environment. We envision a scenario w
herein the human communicates with an AI agent powering such a device by asking

questions (e.g., "where did you last see my keys?"). In order to succeed at this task, the egocentric AI assistant must (1) construct semantically rich and efficient scene memories that encode spatio-temporal information about objects seen during the tour and (2) possess the ability to understand the question and ground its answer into the semantic memory representation. Towards that end, we introduce (1) a new task -- Episodic Memory Question Answering (EMQA) wherein an egocentric AI assistant is provided with a video sequence (the tour) and a question as an input and is asked to localize its answer to the question within the tour, (2) a dataset of grounded questions designed to probe the agent's spatio-temporal understanding of the tour, and (3) a model for the task that encodes the scene as an allocentric, top-down semantic feature map and grounds the question into the map to localize the answer. We show that our choice of episodic scene memory outperforms naive, off-the-shelf solutions for the task as well as a host of very competitive baselines and is robust to noise in depth, pose as well as camera jitter.

****

Continual Stereo Matching of Continuous Driving Scenes With Growing Architecture

Chenghao Zhang, Kun Tian, Bin Fan, Gaofeng Meng, Zhaoxiang Zhang, Chunhong Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18901-18910

The deep stereo models have achieved state-of-the-art performance on driving scenes, but they suffer from severe performance degradation when tested on unseen scenes. Although recent work has narrowed this performance gap through continuous online adaptation, this setup requires continuous gradient updates at inference and can hardly deal with rapidly changing scenes. To address these challenges, we propose to perform continual stereo matching where a model is tasked to 1) continually learn new scenes, 2) overcome forgetting previously learned scenes, and 3) continuously predict disparities at deployment. We achieve this goal by introducing a Reusable Architecture Growth (RAG) framework. RAG leverages task-specific neural unit search and architecture growth for continual learning of new scenes. During growth, it can maintain high reusability by reusing previous neural units while achieving good performance. A module named Scene Router is further introduced to adaptively select the scene-specific architecture path at inference. Experimental results demonstrate that our method achieves compelling performance in various types of challenging driving scenes.

****

Few-Shot Backdoor Defense Using Shapley Estimation

Jiyang Guan, Zhuozhuo Tu, Ran He, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13358-13367

Deep neural networks have achieved impressive performance in a variety of tasks over the last decade, such as autonomous driving, face recognition, and medical diagnosis. However, prior works show that deep neural networks are easily manipulated into specific, attacker-decided behaviors in the inference stage by backdoor attacks which inject malicious small hidden triggers into model training, raising serious security threats. To determine the triggered neurons and protect against backdoor attacks, we exploit Shapley value and develop a new approach called Shapley Pruning (ShapPruning) that successfully mitigates backdoor attacks from models in a data-insufficient situation (1 image per class or even free of data). Considering the interaction between neurons, ShapPruning identifies the few infected neurons (under 1% of all neurons) and manages to protect the model's structure and accuracy after pruning as many infected neurons as possible. To accelerate ShapPruning, we further propose discarding threshold and epsilon-greedy strategy to accelerate Shapley estimation, making it possible to repair poisoned models with only several minutes. Experiments demonstrate the effectiveness and robustness of our method against various attacks and tasks compared to existing methods.

****

Cycle-Consistent Counterfactuals by Latent Transformations

Saeed Khorram, Li Fuxin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10203-10212

CounterFactual (CF) visual explanations try to find images similar to the query image that change the decision of a vision system to a specified outcome. Existing methods either require inference-time optimization or joint training with a generative adversarial model which makes them time-consuming and difficult to use in practice. We propose a novel approach, Cycle-Consistent Counterfactuals by Latent Transformations (C3LT), which learns a latent transformation that automatically generates visual CFs by steering in the latent space of generative models. Our method uses cycle consistency between the query and CF latent representations which helps our training to find better solutions. C3LT can be easily plugged into any state-of-the-art pretrained generative network. This enables our method to generate high-quality and interpretable CF images at high resolution such as those in ImageNet. In addition to several established metrics for evaluating CF explanations, we introduce a novel metric tailored to assess the quality of the generated CF examples and validate the effectiveness of our method on an extensive set of experiments.

****

ADeLA: Automatic Dense Labeling With Attention for Viewpoint Shift in Semantic Segmentation

Hanxiang Ren, Yanchao Yang, He Wang, Bokui Shen, Qingnan Fan, Youyi Zheng, C. Karen Liu, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8079-8089

We describe a method to deal with performance drop in semantic segmentation caused by viewpoint changes within multi-camera systems, where temporally paired images are readily available, but the annotations may only be abundant for a few typical views. Existing methods alleviate performance drop via domain alignment in a shared space and assume that the mapping from the aligned space to the output is transferable. However, the novel content induced by viewpoint changes may nullify such a space for effective alignments, thus resulting in negative adaptation. Our method works without aligning any statistics of the images between the two domains. Instead, it utilizes a novel attention-based view transformation network trained only on color images to hallucinate the semantic images for the target. Despite the lack of supervision, the view transformation network can still generalize to semantic images thanks to the induced "information transport" bias. Furthermore, to resolve ambiguities in converting the semantic images to semantic labels, we treat the view transformation network as a functional representation of an unknown mapping implied by the color images and propose functional label hallucination to generate pseudo-labels with uncertainties in the target domains. Our method surpasses baselines built on state-of-the-art correspondence estimation and view synthesis methods. Moreover, it outperforms the state-of-the-art unsupervised domain adaptation methods that utilize self-training and adversarial domain alignments. Our code and dataset will be made publicly available.

****

Joint Hand Motion and Interaction Hotspots Prediction From Egocentric Videos

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3282-3292

We propose to forecast future hand-object interactions given an egocentric video. Instead of predicting action labels or pixels, we directly predict the hand motion trajectory and the future contact points on the next active object (i.e., interaction hotspots). This relatively low-dimensional representation provides a concrete description of future interactions. To tackle this task, we first provide an automatic way to collect trajectory and hotspots labels in large-scale data. We then use this data to train an Object-Centric Transformer (OCT) model for prediction. Our model performs hand and object interaction reasoning via the self-attention mechanism in Transformers. OCT also provides a probabilistic framework to sample the future trajectory and hotspots to handle uncertainty in prediction. We perform experiments on the Epic-Kitchens-55, Epic-Kitchens-100, and EGTEA Gaze+ datasets, and show that OCT significantly outperforms state-of-the-art approaches by a large margin.

****

Blind Face Restoration via Integrating Face Shape and Generative Priors
Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, Yi ng Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 7662-7671
Blind face restoration, which aims to reconstruct high-quality images from low-q uality inputs, can benefit many applications. Although existing generative-based methods achieve significant progress in producing high-quality images, they oft en fail to restore natural face shapes and high-fidelity facial details from sev erely-degraded inputs. In this work, we propose to integrate shape and generativ e priors to guide the challenging blind face restoration. Firstly, we set up a s hape restoration module to recover reasonable facial geometry with 3D reconstruc tion. Secondly, a pretrained facial generator is adopted as decoder to generate photo-realistic high-resolution images. To ensure high-fidelity, hierarchical sp atial features extracted from the low-quality inputs and rendered 3D images are inserted into the decoder with our proposed Adaptive Feature Fusion Block (AFFB) . Moreover, we introduce hybrid-level losses to jointly train the shape and gene rative priors together with other network parts such that these two priors bette r adapt to our blind face restoration task. The proposed Shape and Generative Pr ior integrated Network (SGPN) can restore high-quality images with clear face sh apes and realistic facial details. Experimental results on synthetic and real-wo rld datasets demonstrate SGPN performs favorably against state-of-the-art blind face restoration methods.
****
MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Vi deo
Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, Junsong Yuan; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13232-13242
Recent transformer-based solutions have been introduced to estimate 3D human pos e from 2D keypoint sequence by considering body joints among all frames globally to learn spatio-temporal correlation. We observe that the motions of different joints differ significantly. However, the previous methods cannot efficiently mo del the solid inter-frame correspondence of each joint, leading to insufficient learning of spatial-temporal correlation. We propose MixSTE (Mixed Spatio-Tempor al Encoder), which has a temporal transformer block to separately model the temp oral motion of each joint and a spatial transformer block to learn inter-joint s patial correlation. These two blocks are utilized alternately to obtain better s patio-temporal feature encoding. In addition, the network output is extended fro m the central frame to entire frames of the input video, thereby improving the c oherence between the input and output sequences. Extensive experiments are condu cted on three benchmarks (Human3.6M, MPI-INF-3DHP, and HumanEva). The results sh ow that our model outperforms the state-of-the-art approach by 10.9% P-MPJPE and 7.6% MPJPE. The code is available at https://github.com/JinluZhang1126/MixSTE.
****
Safe-Student for Safe Deep Semi-Supervised Learning With Unseen-Class Unlabeled Data
Rundong He, Zhongyi Han, Xiankai Lu, Yilong Yin; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14585-14594
Deep semi-supervised learning (SSL) methods aim to take advantage of abundant un labeled data to improve the algorithm performance. In this paper, we consider th e problem of safe SSL scenario where unseen-class instances appear in the unlabe led data. This setting is essential and commonly appears in a variety of real ap plications. One intuitive solution is removing these unseen-class instances afte r detecting them during the SSL process. Nevertheless, the performance of unseen -class identification is limited by the small number of labeled data and ignorin g the availability of unlabeled data. To take advantage of these unseen-class da ta and ensure performance, we propose a safe SSL method called SAFE-STUDENT from the teacher-student view. Firstly, a new scoring function called energy-discrep ancy (ED) is proposed to help the teacher model improve the security of instance s selection. Then, a novel unseen-class label distribution learning mechanism mi

tigates the unseen-class perturbation by calibrating the unseen-class label dist ribution. Finally, we propose an iterative optimization strategy to facilitate t eacher-student network learning. Extensive studies on several representative dat asets show that SAFE-STUDENT remarkably outperforms the state-of-the-art, verify ing the feasibility and robustness of our method in the under-explored problem.
****

Learning To Zoom Inside Camera Imaging Pipeline
Chengzhou Tang, Yuqiang Yang, Bing Zeng, Ping Tan, Shuaicheng Liu; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 2, pp. 17552-17561
Existing single image super-resolution methods are either designed for synthetic data, or for real data but in the RGB-to-RGB or the RAW-to-RGB domain. This pap er proposes to zoom an image from RAW to RAW inside the camera imaging pipeline. The RAW-to-RAW domain closes the gap between the ideal and the real degradation models. It also excludes the image signal processing pipeline, which refocuses the model learning onto the super-resolution. To these ends, we design a method that receives a low-resolution RAW as the input and estimates the desired higher -resolution RAW jointly with the degradation model. In our method, two convoluti onal neural networks are learned to constrain the high-resolution image and the degradation model in lower-dimensional subspaces. This subspace constraint conve rts the ill-posed SISR problem to a well-posed one. To demonstrate the superiori ty of the proposed method and the RAW-to-RAW domain, we conduct evaluations on t he RealSR and the SR-RAW datasets. The results show that our method performs sup eriorly over the state-of-the-arts both qualitatively and quantitatively, and it also generalizes well and enables zero-shot transfer across different sensors.
****

High-Fidelity GAN Inversion for Image Attribute Editing
Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, Qifeng Chen; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11379-11388
We present a novel high-fidelity generative adversarial network (GAN) inversion framework that enables attribute editing with image-specific details well-preser ved (e.g., background, appearance, and illumination). We first analyze the chall enges of high-fidelity GAN inversion from the perspective of lossy data compress ion. With a low bit-rate latent code, previous works have difficulties in preser ving high-fidelity details in reconstructed and edited images. Increasing the si ze of a latent code can improve the accuracy of GAN inversion but at the cost of inferior editability. To improve image fidelity without compromising editabilit y, we propose a distortion consultation approach that employs a distortion map a s a reference for high-fidelity reconstruction. In the distortion consultation i nversion (DCI), the distortion map is first projected to a high-rate latent map, which then complements the basic low-rate latent code with more details via con sultation fusion. To achieve high-fidelity editing, we propose an adaptive disto rtion alignment (ADA) module with a self-supervised training scheme, which bridg es the gap between the edited and inversion images. Extensive experiments in the face and car domains show a clear improvement in both inversion and editing qua lity.
****

RCP: Recurrent Closest Point for Point Cloud
Xiaodong Gu, Chengzhou Tang, Weihao Yuan, Zuozhuo Dai, Siyu Zhu, Ping Tan; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 8216-8226
3D motion estimation including scene flow and point cloud registration has drawn increasing interest. Inspired by 2D flow estimation, recent methods employ deep neural networks to construct the cost volume for estimating accurate 3D flow. H owever, these methods are limited by the fact that it is difficult to define a s earch window on point clouds because of the irregular data structure. In this pa per, we avoid this irregularity by a simple yet effective method. We decompose t he problem into two interlaced stages, where the 3D flows are optimized point-wi sely at the first stage and then globally regularized in a recurrent network at

the second stage. Therefore, the recurrent network only receives the regular poi nt-wise information as the input. In the experiments, we evaluate the proposed m ethod on both the 3D scene flow estimation and the point cloud registration task . For 3D scene flow estimation, we make comparisons on the widely used FlyingThi ngs3D and KITTI datasets. For point cloud registration, we follow previous works and evaluate the data pairs with large pose and partially overlapping from Mode lNet40. The results show that our method outperforms the previous method and ach ieves a new state-of-the-art performance on both 3D scene flow estimation and po int cloud registration, which demonstrates the superiority of the proposed zero- order method on irregular point cloud data. Our source code is available at http s://github.com/gxd1994/RCP.
****

gDNA: Towards Generative Detailed Neural Avatars
Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geige r, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20427-20437
To make 3D human avatars widely available, we must be able to generate a variety of 3D virtual humans with varied identities and shapes in arbitrary poses. This task is challenging due to the diversity of clothed body shapes, their complex articulations, and the resulting rich, yet stochastic geometric detail in clothi ng. Hence, current methods to represent 3D people do not provide a full generati ve model of people in clothing. In this paper, we propose a novel method that le arns to generate detailed 3D shapes of people in a variety of garments with corr esponding skinning weights. Specifically, we devise a multi-subject forward skin ning module that is learned from only a few posed, un-rigged scans per subject. To capture the stochastic nature of high-frequency details in garments, we lever age an adversarial loss formulation that encourages the model to capture the und erlying statistics. We provide empirical evidence that this leads to realistic g eneration of local details such as wrinkles. We show that our model is able to g enerate natural human avatars wearing diverse and detailed clothing. Furthermore , we show that our method can be used on the task of fitting human models to raw scans, outperforming the previous state-of-the-art.
****

A Dual Weighting Label Assignment Scheme for Object Detection
Shuai Li, Chenhang He, Ruihuang Li, Lei Zhang; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9387-9396
Label assignment (LA), which aims to assign each training sample a positive (pos) and a negative (neg) loss weight, plays an important role in object detection. Existing LA methods mostly focus on the design of pos weighting function, while the neg weight is directly derived from the pos weight. Such a mechanism limits the learning capacity of detectors. In this paper, we explore a new weighting p aradigm, termed dual weighting (DW), to specify pos and neg weights separately. We first identify the key influential factors of pos/neg weights by analyzing th e evaluation metrics in object detection, and then design the pos and neg weight ing functions based on them. Specifically, the pos weight of a sample is determi ned by the consistency degree between its classification and localization scores , while the neg weight is decomposed into two terms: the probability that it is a neg sample and its importance conditioned on being a neg sample. Such a weight ing strategy offers greater flexibility to distinguish between important and les s important samples, resulting in a more effective object detector. Equipped wit h the proposed DW method, a single FCOS-ResNet-50 detector can reach 41.5% mAP o n COCO under 1xschedule, outperforming other existing LA methods. It consistentl y improves the baselines on COCO by a large margin under various backbones witho ut bells and whistles. Code is available at https://github.com/strongwolf/DW.
****

FAM: Visual Explanations for the Feature Representations From Deep Convolutional Networks
Yuxi Wu, Changhuai Chen, Jun Che, Shiliang Pu; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10307-10316
In recent years, increasing attention has been drawn to the internal mechanisms

of representation models. Traditional methods are inapplicable to fully explain the feature representations, especially if the images do not fit into any category. In this case, employing an existing class or the similarity with other image is unable to provide a complete and reliable visual explanation. To handle this task, we propose a novel visual explanation paradigm called Feature Activation Mapping (FAM) in this paper. Under this paradigm, Grad-FAM and Score-FAM are designed for visualizing feature representations. Unlike the previous approaches, FAM locates the regions of images that contribute most to the feature vector itself. Extensive experiments and evaluations, both subjective and objective, showed that Score-FAM provided most promising interpretable visual explanations for feature representations in Person Re-Identification. Furthermore, FAM also can be employed to analyze other vision tasks, such as self-supervised representation learning.

****

Hyperbolic Vision Transformers: Combining Improvements in Metric Learning

Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, Ivan Oseledets; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7409-7419

Metric learning aims to learn a highly discriminative model encouraging the embeddings of similar classes to be close in the chosen metrics and pushed apart for dissimilar ones. The common recipe is to use an encoder to extract embeddings and a distance-based loss function to match the representations -- usually, the Euclidean distance is utilized. An emerging interest in learning hyperbolic data embeddings suggests that hyperbolic geometry can be beneficial for natural data. Following this line of work, we propose a new hyperbolic-based model for metric learning. At the core of our method is a vision transformer with output embeddings mapped to hyperbolic space. These embeddings are directly optimized using modified pairwise cross-entropy loss. We evaluate the proposed model with six different formulations on four datasets achieving the new state-of-the-art performance. The source code is available at https://github.com/htdt/hyp_metric.

****

MaskGIT: Masked Generative Image Transformer

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, William T. Freeman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11315-11325

Generative transformers have experienced rapid popularity growth in the computer vision community in synthesizing high-fidelity and high-resolution images. The best generative transformer models so far, however, still treat an image naively as a sequence of tokens, and decode an image sequentially following the raster scan ordering (i.e. line-by-line). We find this strategy neither optimal nor efficient. This paper proposes a novel image synthesis paradigm using a bidirectional transformer decoder, which we term MaskGIT. During training, MaskGIT learns to predict randomly masked tokens by attending to tokens in all directions. At inference time, the model begins with generating all tokens of an image simultaneously, and then refines the image iteratively conditioned on the previous generation. Our experiments demonstrate that MaskGIT significantly outperforms the state-of-the-art transformer model on the ImageNet dataset, and accelerates autoregressive decoding by up to 48x. Besides, we illustrate that MaskGIT can be easily extended to various image editing tasks, such as inpainting, extrapolation, and image manipulation. Project page: masked-generative-image-transformer.github.io.

****

Revisiting the "Video" in Video-Language Understanding

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2917-2927

What makes a video task uniquely suited for videos, beyond what can be understood from a single image? Building on recent progress in self-supervised image-language models, we revisit this question in the context of video and language tasks. We propose the atemporal probe (ATP), a new model for video-language analysis which provides a stronger bound on the baseline accuracy of multimodal models co

nstrained by image-level understanding. By applying this model to standard discr iminative video and language tasks, such as video question answering and text-to -video retrieval, we characterize the limitations and potential of current video -language benchmarks. We find that understanding of event temporality is often n ot necessary to achieve strong or state-of-the-art performance, even compared wi th recent large-scale video-language models and in contexts intended to benchmar k deeper video-level understanding. We also demonstrate how ATP can improve both video-language dataset and model design. We describe a technique for leveraging ATP to better disentangle dataset subsets with a higher concentration of tempor ally challenging data, improving benchmarking efficacy for causal and temporal u nderstanding. Further, we show that effectively integrating ATP into full video- level temporal models can improve efficiency and state-of-the-art accuracy.
****

Local Texture Estimator for Implicit Representation Function
Jaewon Lee, Kyong Hwan Jin; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 1929-1938
Recent works with an implicit neural function shed light on representing images in arbitrary resolution. However, a standalone multi-layer perceptron shows limi ted performance in learning high-frequency components. In this paper, we propose a Local Texture Estimator (LTE), a dominant-frequency estimator for natural ima ges, enabling an implicit function to capture fine details while reconstructing images in a continuous manner. When jointly trained with a deep super-resolution (SR) architecture, LTE is capable of characterizing image textures in 2D Fourie r space. We show that an LTE-based neural function achieves favorable performanc e against existing deep SR methods within an arbitrary-scale factor. Furthermore , we demonstrate that our implementation takes the shortest running time compare d to previous works.
****

Instance-Aware Dynamic Neural Network Quantization
Zhenhua Liu, Yunhe Wang, Kai Han, Siwei Ma, Wen Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12434-1 2443
Quantization is an effective way to reduce the memory and computational costs of deep neural networks in which the full-precision weights and activations are re presented using low-bit values. The bit-width for each layer in most of existing quantization methods is static, i.e., the same for all samples in the given dat aset. However, natural images are of huge diversity with abundant content and us ing such a universal quantization configuration for all samples is not an optima l strategy. In this paper, we present to conduct the low-bit quantization for ea ch image individually, and develop a dynamic quantization scheme for exploring t heir optimal bit-widths. To this end, a lightweight bit-controller is establishe d and trained jointly with the given neural network to be quantized. During infe rence, the quantization configuration for an arbitrary image will be determined by the bit-widths generated by the controller, e.g., an image with simple textur e will be allocated with lower bits and computational complexity and vice versa. Experimental results conducted on benchmarks demonstrate the effectiveness of t he proposed dynamic quantization method for achieving state-of-art performance i n terms of accuracy and computational complexity. The code will be available at https://github.com/huawei-noah/Efficient-Computing and https://gitee.com/mindspo re/models/tree/master/research/cv/DynamicQuant.
****

When To Prune? A Policy Towards Early Structural Pruning
Maying Shen, Pavlo Molchanov, Hongxu Yin, Jose M. Alvarez; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1 2247-12256
Pruning enables appealing reductions in network memory footprint and time comple xity. Conventional post-training pruning techniques lean towards efficient infer ence while overlooking the heavy computation for training. Recent exploration of pre-training pruning at initialization hints on training cost reduction via pru ning, but suffers noticeable performance degradation. We attempt to combine the

benefits of both directions and propose a policy that prunes as early as possibl e during training without hurting performance. Instead of pruning at initializat ion, our method exploits initial dense training for few epochs to quickly guide the architecture, while constantly evaluating dominant sub-networks via neuron i mportance ranking. This unveils dominant sub-networks whose structures turn stab le, allowing conventional pruning to be pushed earlier into the training. To do this early, we further introduce an Early Pruning Indicator (EPI) that relies on sub-network architectural similarity and quickly triggers pruning when the sub- network's architecture stabilizes. Through extensive experiments on ImageNet, we show that EPI empowers a quick tracking of early training epochs suitable for p runing, offering same efficacy as an otherwise "oracle" grid-search that scans t hrough epochs and requires orders of magnitude more compute. Our method yields 1 .4% top-1 accuracy boost over state-of-the-art pruning counterparts, cuts down t raining cost on GPU by 2.4x, hence offers a new efficiency-accuracy boundary for network pruning during training.

****

COTS: Collaborative Two-Stream Vision-Language Pre-Training Model for Cross-Moda l Retrieval

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, Ji-Rong Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 15692-15701

Large-scale single-stream pre-training has shown dramatic performance in image-t ext retrieval. Regrettably, it faces low inference efficiency due to heavy atten tion layers. Recently, two-stream methods like CLIP and ALIGN with high inferenc e efficiency have also shown promising performance, however, they only consider instance-level alignment between the two streams (thus there is still room for i mprovement). To overcome these limitations, we propose a novel COllaborative Two -Stream vision-language pre-training model termed COTS for image-text retrieval by enhancing cross-modal interaction. In addition to instance-level alignment vi a momentum contrastive learning, we leverage two extra levels of cross-modal int eractions in our COTS: (1) Token-level interaction -- a masked vision-language m odeling (MVLM) learning objective is devised without using a cross-stream networ k module, where variational autoencoder is imposed on the visual encoder to gene rate visual tokens for each image. (2) Task-level interaction -- a KL-alignment learning objective is devised between text-to-image and image-to-text retrieval tasks, where the probability distribution per task is computed with the negative queues in momentum contrastive learning. Under a fair comparison setting, our C OTS achieves the highest performance among all two-stream methods and comparable performance (but with 10,800x faster in inference) w.r.t. the latest single-str eam methods. Importantly, our COTS is also applicable to text-to-video retrieval , yielding new state-of-the-art on the widely-used MSR-VTT dataset.

****

Degree-of-Linear-Polarization-Based Color Constancy

Taishi Ono, Yuhi Kondo, Legong Sun, Teppei Kurita, Yusuke Moriuchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 19740-19749

Color constancy is an essential function in digital photography and a fundamenta l process for many computer vision applications. Accordingly, many methods have been proposed, and some recent ones have used deep neural networks to handle mor e complex scenarios. However, both the traditional and latest methods still impo se strict assumptions on their target scenes in explicit or implicit ways. This paper shows that the degree of linear polarization dramatically solves the color constancy problem because it allows us to find achromatic pixels stably. Becaus e we only rely on the physics-based polarization model, we significantly reduce the assumptions compared to existing methods. Furthermore, we captured a wide va riety of scenes with ground-truth illuminations for evaluation, and the proposed approach achieved state-of-the-art performance with a low computational cost. A dditionally, the proposed method can estimate illumination colors from chromatic pixels and manage multi-illumination scenes. Lastly, the evaluation scenes and codes are publicly available to encourage more development in this field.

****
A Voxel Graph CNN for Object Classification With Event Cameras
Yongjian Deng, Hao Chen, Hai Liu, Youfu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1172-1181
Event cameras attract researchers' attention due to their low power consumption, high dynamic range, and extremely high temporal resolution. Learning models on event-based object classification have recently achieved massive success by accumulating sparse events into dense frames to apply traditional 2D learning methods. Yet, these approaches necessitate heavy-weight models and are with high computational complexity due to the redundant information introduced by the sparse-to-dense conversion, limiting the potential of event cameras on real-life applications. This study aims to address the core problem of balancing accuracy and model complexity for event-based classification models. To this end, we introduce a novel graph representation for event data to exploit their sparsity better and customize a lightweight voxel graph convolutional neural network (EV-VGCNN) for event-based classification. Specifically, (1) using voxel-wise vertices rather than previous point-wise inputs to explicitly exploit regional 2D semantics of event streams while keeping the sparsity; (2) proposing a multi-scale feature relational layer (MFRL) to extract spatial and motion cues from each vertex discriminatively concerning its distances to neighbors. Comprehensive experiments show that our model can advance state-of-the-art classification accuracy with extremely low model complexity (merely 0.84M parameters).
****
On the Importance of Asymmetry for Siamese Representation Learning
Xiao Wang, Haoqi Fan, Yuandong Tian, Daisuke Kihara, Xinlei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16570-16579
Many recent self-supervised frameworks for visual representation learning are based on certain forms of Siamese networks. Such networks are conceptually symmetric with two parallel encoders, but often practically asymmetric as numerous mechanisms are devised to break the symmetry. In this work, we conduct a formal study on the importance of asymmetry by explicitly distinguishing the two encoders within the network -- one produces source encodings and the other targets. Our key insight is keeping a relatively lower variance in target than source generally benefits learning. This is empirically justified by our results from five case studies covering different variance-oriented designs, and is aligned with our preliminary theoretical analysis on the baseline. Moreover, we find the improvements from asymmetric designs generalize well to longer training schedules, multiple other frameworks and newer backbones. Finally, the combined effect of several asymmetric designs achieves a state-of-the-art accuracy on ImageNet linear probing and competitive results on downstream transfer. We hope our exploration will inspire more research in exploiting asymmetry for Siamese representation learning.
****
Probing Representation Forgetting in Supervised and Unsupervised Continual Learning
MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, Eugene Belilovsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16712-16721
Continual Learning (CL) research typically focuses on tackling the phenomenon of catastrophic forgetting in neural networks. Catastrophic forgetting is associated with an abrupt loss of knowledge previously learned by a model when the task, or more broadly the data distribution, being trained on changes. In supervised learning problems this forgetting, resulting from a change in the model's representation, is typically measured or observed by evaluating the decrease in old task performance. However, a model's representation can change without losing knowledge about prior tasks. In this work we consider the concept of representation forgetting, observed by using the difference in performance of an optimal linear classifier before and after a new task is introduced. Using this tool we revisit a number of standard continual learning benchmarks and observe that, through t

his lens, model representations trained without any explicit control for forgett ing often experience small representation forgetting and can sometimes be compar able to methods which explicitly control for forgetting, especially in longer ta sk sequences. We also show that representation forgetting can lead to new insigh ts on the effect of model capacity and loss function used in continual learning. Based on our results, we show that a simple yet competitive approach is to lear n representations continually with standard supervised contrastive learning whil e constructing prototypes of class samples when queried on old samples.
****

ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval
Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, Jingdong Wang; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 51 84-5193
Visual appearance is considered to be the most important cue to understand image s for cross-modal retrieval, while sometimes the scene text appearing in images can provide valuable information to understand the visual semantics. Most of exi sting cross-modal retrieval approaches ignore the usage of scene text informatio n and directly adding this information may lead to performance degradation in sc ene text free scenarios. To address this issue, we propose a full transformer ar chitecture to unify these cross-modal retrieval scenarios in a single Vision and Scene Text Aggregation framework (ViSTA). Specifically, ViSTA utilizes transfor mer blocks to directly encode image patches and fuse scene text embedding to lea rn an aggregated visual representation for cross-modal retrieval. To tackle the modality missing problem of scene text, we propose a novel fusion token based tr ansformer aggregation approach to exchange the necessary scene text information only through the fusion token and concentrate on the most important features in each modality. To further strengthen the visual modality, we develop dual contra stive learning losses to embed both image-text pairs and fusion-text pairs into a common cross-modal space. Compared to existing methods, ViSTA enables to aggre gate relevant scene text semantics with visual appearance, and hence improve res ults under both scene text free and scene text aware scenarios. Experimental res ults show that ViSTA outperforms other methods by at least 8.4% at Recall@1 for scene text aware retrieval task. Compared with state-of-the-art scene text free retrieval methods, ViSTA can achieve better accuracy on Flicker30K and MSCOCO wh ile running at least three times faster during the inference stage, which valida tes the effectiveness of the proposed framework.
****

DenseCLIP: Language-Guided Dense Prediction With Context-Aware Prompting
Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 18082-18091
Recent progress has shown that large-scale pre-training using contrastive image-text pairs can be a promising alternative for high-quality visual representation learning from natural language supervision. Benefiting from a broader source of supervision, this new paradigm exhibits impressive transferability to downstrea m classification tasks and datasets. However, the problem of transferring the kn owledge learned from image-text pairs to more complex dense prediction tasks has barely been visited. In this work, we present a new framework for dense predict ion by implicitly and explicitly leveraging the pre-trained knowledge from CLIP. To this end, we convert the original image-text matching problem in CLIP to a p ixel-text matching problem and use the pixel-text score maps to guide the learni ng of dense prediction models. By further using the contextual information from the image to prompt the language model, we are able to facilitate our model to b etter exploit the pre-trained knowledge. Our method is model-agnostic, which can be applied to arbitrary dense prediction systems and various pre-trained visual backbones including both CLIP models and ImageNet pre-trained models. Extensive experiments demonstrate the superior performance of our methods on semantic seg mentation, object detection, and instance segmentation tasks.
****

Exploring Effective Data for Surrogate Training Towards Black-Box Attack
Xuxiang Sun, Gong Cheng, Hongda Li, Lei Pei, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15355-15364
Without access to the training data where a black-box victim model is deployed, training a surrogate model for black-box adversarial attack is still a struggle. In terms of data, we mainly identify three key measures for effective surrogate training in this paper. First, we show that leveraging the loss introduced in this paper to enlarge the inter-class similarity makes more sense than enlarging the inter-class diversity like existing methods. Next, unlike the approaches that expand the intra-class diversity in an implicit model-agnostic fashion, we propose a loss function specific to the surrogate model for our generator to enhance the intra-class diversity. Finally, in accordance with the in-depth observations for the methods based on proxy data, we argue that leveraging the proxy data is still an effective way for surrogate training. To this end, we propose a triple-player framework by introducing a discriminator into the traditional data-free framework. In this way, our method can be competitive when there are few semantic overlaps between the scarce proxy data (with the size between 1k and 5k) and the training data. We evaluate our method on a range of victim models and datasets. The extensive results witness the effectiveness of our method. Our source code is available at https://github.com/xuxiangsun/ST-Data.
****
JRDB-Act: A Large-Scale Dataset for Spatio-Temporal Action, Social Group and Activity Detection
Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, Hamid Rezatofighi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20983-20992
The availability of large-scale video action understanding datasets has facilitated advances in the interpretation of visual scenes containing people. However, learning to recognise human actions and their social interactions in an unconstrained real-world environment comprising numerous people, with potentially highly unbalanced and long-tailed distributed action labels from a stream of sensory data captured from a mobile robot platform remains a significant challenge, not least owing to the lack of a reflective large-scale dataset. In this paper, we introduce JRDB-Act, as an extension of the existing JRDB, which is captured by a social mobile manipulator and reflects a real distribution of human daily-life actions in a university campus environment. JRDB-Act has been densely annotated with atomic actions, comprises over 2.8M action labels, constituting a large-scale spatio-temporal action detection dataset. Each human bounding box is labeled with one pose-based action label and multiple (optional) interaction-based action labels. Moreover JRDB-Act provides social group annotation, conducive to the task of grouping individuals based on their interactions in the scene to infer their social activities (common activities in each social group). Each annotated label in JRDB-Act is tagged with the annotators' confidence level which contributes to the development of reliable evaluation strategies. In order to demonstrate how one can effectively utilise such annotations, we develop an end-to-end trainable pipeline to learn and infer these tasks, i.e. individual action and social group detection. The data and the evaluation code will be publicly available at https://jrdb.erc.monash.edu/.
****
AR-NeRF: Unsupervised Learning of Depth and Defocus Effects From Natural Images With Aperture Rendering Neural Radiance Fields
Takuhiro Kaneko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18387-18397
Fully unsupervised 3D representation learning has gained attention owing to its advantages in data collection. A successful approach involves a viewpoint-aware approach that learns an image distribution based on generative models (e.g., generative adversarial networks (GANs)) while generating various view images based on 3D-aware models (e.g., neural radiance fields (NeRFs)). However, they require images with various views for training, and consequently, their application to

datasets with few or limited viewpoints remains a challenge. As a complementary approach, an aperture rendering GAN (AR-GAN) that employs a defocus cue was proposed. However, an AR-GAN is a CNN-based model and represents a defocus independently from a viewpoint change despite its high correlation, which is one of the reasons for its performance. As an alternative to an AR-GAN, we propose an aperture rendering NeRF (AR-NeRF), which can utilize viewpoint and defocus cues in a unified manner by representing both factors in a common ray-tracing framework. Moreover, to learn defocus-aware and defocus-independent representations in a disentangled manner, we propose aperture randomized training, for which we learn to generate images while randomizing the aperture size and latent codes independently. During our experiments, we applied AR-NeRF to various natural image datasets, including flower, bird, and face images, the results of which demonstrate the utility of AR-NeRF for unsupervised learning of the depth and defocus effects.
****

Likert Scoring With Grade Decoupling for Long-Term Action Assessment
Angchi Xu, Ling-An Zeng, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3232-3241
Long-term action quality assessment is a task of evaluating how well an action is performed, namely, estimating a quality score from a long video. Intuitively, longterm actions generally involve parts exhibiting different levels of skill, and we call the levels of skill as performance grades. For example, technical highlights and faults may appear in the same long-term action. Hence, the final score should be determined by the comprehensive effect of different grades exhibited in the video. To explore this latent relationship, we design a novel Likert scoring paradigm inspired by the Likert scale in psychometrics, in which we quantify the grades explicitly and generate the final quality score by combining the quantitative values and the corresponding responses estimated from the video, instead of performing direct regression. Moreover, we extract gradespecific features, which will be used to estimate the responses of each grade, through a Transformer decoder architecture with diverse learnable queries. The whole model is named as Grade-decoupling Likert Transformer (GDLT), and we achieve state-of-the-art results on two long-term action assessment datasets.
****

Many-to-Many Splatting for Efficient Video Frame Interpolation
Ping Hu, Simon Niklaus, Stan Sclaroff, Kate Saenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3553-3562
Motion-based video frame interpolation commonly relies on optical flow to warp pixels from the inputs to the desired interpolation instant. Yet due to the inherent challenges of motion estimation (e.g. occlusions and discontinuities), most state-of-the-art interpolation approaches require subsequent refinement of the warped result to generate satisfying outputs, which drastically decreases the efficiency for multi-frame interpolation. In this work, we propose a fully differentiable Many-to-Many (M2M) splatting framework to interpolate frames efficiently. Specifically, given a frame pair, we estimate multiple bidirectional flows to directly forward warp the pixels to the desired time step, and then fuse any overlapping pixels. In doing so, each source pixel renders multiple target pixels and each target pixel can be synthesized from a larger area of visual context. This establishes a many-to-many splatting scheme with robustness to artifacts like holes. Moreover, for each input frame pair, M2M only performs motion estimation once and has a minuscule computational overhead when interpolating an arbitrary number of in-between frames, hence achieving fast multi-frame interpolation. We conducted extensive experiments to analyze M2M, and found that it significantly improves the efficiency while maintaining high effectiveness.
****

Investigating Top-k White-Box and Transferable Black-Box Attack
Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15085-15094
Existing works have identified the limitation of top-1 attack success rate (ASR)

as a metric to evaluate the attack strength but exclusively investigated it in the white-box setting, while our work extends it to a more practical black-box setting: transferable attack. It is widely reported that stronger I-FGSM transfers worse than simple FGSM, leading to a popular belief that transferability is at odds with the white-box attack strength. Our work challenges this belief with empirical finding that stronger attack actually transfers better for the general top-k ASR indicated by the interest class rank (ICR) after attack. For increasing the attack strength, with an intuitive interpretation of the logit gradient from the geometric perspective, we identify that the weakness of the commonly used losses lie in prioritizing the speed to fool the network instead of maximizing its strength. To this end, we propose a new normalized CE loss that guides the logit to be updated in the direction of implicitly maximizing its rank distance from the ground-truth class. Extensive results in various settings have verified that our proposed new loss is simple yet effective for top-k attack. Code is available at: https://bit.ly/3uCiomP

****

Decoupling and Recoupling Spatiotemporal Representation for RGB-D-Based Motion Recognition

Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, Rong Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20154-20163

Decoupling spatiotemporal representation refers to decomposing the spatial and temporal features into dimension-independent factors. Although previous RGB-D-based motion recognition methods have achieved promising performance through the tightly coupled multi-modal spatiotemporal representation, they still suffer from (i) optimization difficulty under small data setting due to the tightly spatiotemporal-entangled modeling; (ii) information redundancy as it usually contains lots of marginal information that is weakly relevant to classification; and (iii) low interaction between multi-modal spatiotemporal information caused by insufficient late fusion. To alleviate these drawbacks, we propose to decouple and recouple spatiotemporal representation for RGB-D-based motion recognition. Specifically, we disentangle the task of learning spatiotemporal representation into 3 sub-tasks: (1) Learning high-quality and dimension independent features through a decoupled spatial and temporal modeling network. (2) Recoupling the decoupled representation to establish stronger space-time dependency. (3) Introducing a Cross-modal Adaptive Posterior Fusion (CAPF) mechanism to capture cross-modal spatiotemporal information from RGB-D data. Seamless combination of these novel designs forms a robust spatiotemporal representation and achieves better performance than state-of-the-art methods on four public motion datasets. Our code is available at https://github.com/damo-cv/MotionRGBD.

****

Learning To Learn by Jointly Optimizing Neural Architecture and Weights

Yadong Ding, Yu Wu, Chengyue Huang, Siliang Tang, Yi Yang, Longhui Wei, Yueting Zhuang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 129-138

Meta-learning enables models to adapt to new environments rapidly with a few training examples. Current gradient-based meta-learning methods concentrate on finding good initialization (meta-weights) for learners but ignore the impact of neural architectures. In this paper, we aim to obtain better meta-learners by co-optimizing the architecture and meta-weights simultaneously. Existing NAS-based meta-learning methods apply a two-stage strategy, i.e., first searching architectures and then re-training meta-weights on the searched architecture. However, this two-stage strategy would break the mutual impact of the architecture and meta-weights since they are optimized separately. Differently, we propose progressive connection consolidation, fixing the architecture layer by layer, in which the layer with the largest weight value would be fixed first. In this way, we can jointly search architectures and train the meta-weights on fixed layers. Besides, to improve the generalization performance of the searched meta-learner on all tasks, we propose a more effective rule for co-optimization, namely Connection-Adaptive Meta-learning (CAML). By searching only once, we can obtain both adaptiv

e architecture and meta-weights for meta-learning. Extensive experiments show th at our method achieves state-of-the-art performance with 3x less computational c ost, revealing our method's effectiveness and efficiency.
****

Attributable Visual Similarity Learning
Borui Zhang, Wenzhao Zheng, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7532-7541
This paper proposes an attributable visual similarity learning (AVSL) framework for a more accurate and explainable similarity measure between images. Most exis ting similarity learning methods exacerbate the unexplainability by mapping each sample to a single point in the embedding space with a distance metric (e.g., M ahalanobis distance, Euclidean distance). Motivated by the human semantic simila rity cognition, we propose a generalized similarity learning paradigm to represe nt the similarity between two images with a graph and then infer the overall sim ilarity accordingly. Furthermore, we establish a bottom-up similarity constructi on and top-down similarity inference framework to infer the similarity based on semantic hierarchy consistency. We first identify unreliable higher-level simila rity nodes and then correct them using the most coherent adjacent lower-level si milarity nodes, which simultaneously preserve traces for similarity attribution. Extensive experiments on the CUB-200-2011, Cars196, and Stanford Online Product s datasets demonstrate significant improvements over existing deep similarity le arning methods and verify the interpretability of our framework.
****

A Self-Supervised Descriptor for Image Copy Detection
Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, Matthijs Douze ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 14532-14542
Image copy detection is an important task for content moderation. We introduce S SCD, a model that builds on a recent self-supervised contrastive training object ive. We adapt this method to the copy detection task by changing the architectur e and training objective, including a pooling operator from the instance matchin g literature, and adapting contrastive learning to augmentations that combine im ages. Our approach relies on an entropy regularization term, promoting consisten t separation between descriptor vectors, and we demonstrate that this significan tly improves copy detection accuracy. Our method produces a compact descriptor v ector, suitable for real-world web scale applications. Statistical information f rom a background image distribution can be incorporated into the descriptor. On the recent DISC2021 benchmark, SSCD is shown to outperform both baseline copy de tection models and self-supervised architectures designed for image classificati on by huge margins, in all settings. For example, SSCD outperforms SimCLR descri ptors by 48% absolute. Code is available at https://github.com/facebookresearch/ sscd-copy-detection.
****

DyTox: Transformers for Continual Learning With DYnamic TOken eXpansion
Arthur Douillard, Alexandre Ramé, Guillaume Couairon, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 022, pp. 9285-9295
Deep network architectures struggle to continually learn new tasks without forge tting the previous tasks. A recent trend indicates that dynamic architectures ba sed on an expansion of the parameters can reduce catastrophic forgetting efficie ntly in continual learning. However, existing approaches often require a task id entifier at test-time, need complex tuning to balance the growing number of para meters, and barely share any information across tasks. As a result, they struggl e to scale to a large number of tasks without significant overhead. In this pape r, we propose a transformer architecture based on a dedicated encoder/decoder fr amework. Critically, the encoder and decoder are shared among all tasks. Through a dynamic expansion of special tokens, we specialize each forward of our decode r network on a task distribution. Our strategy scales to a large number of tasks while having negligible memory and time overheads due to strict control of the parameters expansion. Moreover, this efficient strategy doesn't need any hyperpa

rameter tuning to control the network's expansion. Our model reaches excellent r
esults on CIFAR100 and state-of-the-art performances on the large-scale ImageNet
100 and ImageNet1000 while having less parameters than concurrent dynamic framew
orks.
****

Towards Robust and Adaptive Motion Forecasting: A Causal Representation Perspect
ive

Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, Alexandre Alahi;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2022, pp. 17081-17092

Learning behavioral patterns from observational data has been a de-facto approac
h to motion forecasting. Yet, the current paradigm suffers from two shortcomings
: brittle under distribution shifts and inefficient for knowledge transfer. In t
his work, we propose to address these challenges from a causal representation pe
rspective. We first introduce a causal formalism of motion forecasting, which ca
sts the problem as a dynamic process with three groups of latent variables, name
ly invariant variables, style confounders, and spurious features. We then introd
uce a learning framework that treats each group separately: (i) unlike the commo
n practice mixing datasets collected from different locations, we exploit their
subtle distinctions by means of an invariance loss encouraging the model to supp
ress spurious correlations; (ii) we devise a modular architecture that factorize
s the representations of invariant mechanisms and style confounders to approxima
te a sparse causal graph; (iii) we introduce a style contrastive loss that not o
nly enforces the structure of style representations but also serves as a self-su
pervisory signal for test-time refinement on the fly. Experiments on synthetic a
nd real datasets show that our proposed method improves the robustness and reusa
bility of learned motion representations, significantly outperforming prior stat
e-of-the-art motion forecasting models for out-of-distribution generalization an
d low-shot transfer.
****

Manifold Learning Benefits GANs

Yao Ni, Piotr Koniusz, Richard Hartley, Richard Nock; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11265-
11274

In this paper, we improve Generative Adversarial Networks by incorporating a man
ifold learning step into the discriminator. We consider locality-constrained lin
ear and subspace-based manifolds, and locality-constrained non-linear manifolds.
 In our design, the manifold learning and coding steps are intertwined with laye
rs of the discriminator, with the goal of attracting intermediate feature repres
entations onto manifolds. We adaptively balance the discrepancy between feature
representations and their manifold view, which is a trade-off between denoising
on the manifold and refining the manifold. We find that locality-constrained non
-linear manifolds outperform linear manifolds due to their non-uniform density a
nd smoothness. We also substantially outperform state-of-the-art baselines.
****

A Keypoint-Based Global Association Network for Lane Detection

Jinsheng Wang, Yinchao Ma, Shaofei Huang, Tianrui Hui, Fei Wang, Chen Qian, Tian
zhu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2022, pp. 1392-1401

Lane detection is a challenging task that requires predicting complex topology s
hapes of lane lines and distinguishing different types of lanes simultaneously.
Earlier works follow a top-down roadmap to regress predefined anchors into vario
us shapes of lane lines, which lacks enough flexibility to fit complex shapes of
 lanes due to the fixed anchor shapes. Lately, some works propose to formulate l
ane detection as a keypoint estimation problem to describe the shapes of lane li
nes more flexibly and gradually group adjacent keypoints belonging to the same l
ane line in a point-by-point manner, which is inefficient and time-consuming dur
ing postprocessing. In this paper, we propose a Global Association Network (GANe
t) to formulate the lane detection problem from a new perspective, where each ke
ypoint is directly regressed to the starting point of the lane line instead of p

oint-by-point extension. Concretely, the association of keypoints to their belon ged lane line is conducted by predicting their offsets to the corresponding star ting points of lanes globally without dependence on each other, which could be d one in parallel to greatly improve efficiency. In addition, we further propose a Lane-aware Feature Aggregator (LFA), which adaptively captures the local correl ations between adjacent keypoints to supplement local information to the global association. Extensive experiments on two popular lane detection benchmarks show that our method outperforms previous methods with F1 score of 79.63% on CULane and 97.71% on Tusimple dataset with high FPS.
****

Negative-Aware Attention Framework for Image-Text Matching
Kun Zhang, Zhendong Mao, Quan Wang, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15661-15 670
Image-text matching, as a fundamental task, bridges the gap between vision and l anguage. The key of this task is to accurately measure similarity between these two modalities. Prior work measuring this similarity mainly based on matched fra gments (i.e., word/region with high relevance), while underestimating or even ig noring the effect of mismatched fragments (i.e., word/region with low relevance) , e.g., via a typical LeaklyReLU or ReLU operation that forces negative scores c lose or exact to zero in attention. This work argues that mismatched textual fra gments, which contain rich mismatching clues, are also crucial for image-text ma tching. We thereby propose a novel Negative-Aware Attention Framework (NAAF), wh ich explicitly exploits both the positive effect of matched fragments and the ne gative effect of mismatched fragments to jointly infer image-text similarity. NA AF (1) delicately designs an iterative optimization method to maximally mine the mismatched fragments, facilitating more discriminative and robust negative effe cts, and (2) devises the two-branch matching mechanism to precisely calculate si milarity/dissimilarity degrees for matched/mismatched fragments with different m asks. Extensive experiments on two benchmark datasets, i.e., Flickr30K and MSCOC O, demonstrate the superior effectiveness of our NAAF, achieving state-of-the-ar t performance. Code will be released at: https://github.com/CrossmodalGroup/NAAF .
****

Semantic-Aligned Fusion Transformer for One-Shot Object Detection
Yizhou Zhao, Xun Guo, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7601-7611
One-shot object detection aims at detecting novel objects according to merely on e given instance. With extreme data scarcity, current approaches explore various feature fusions to obtain directly transferable meta-knowledge. Yet, their perf ormances are often unsatisfactory. In this paper, we attribute this to inappropr iate correlation methods that misalign query-support semantics by overlooking sp atial structures and scale variances. Upon analysis, we leverage the attention m echanism and propose a simple but effective architecture named Semantic-aligned Fusion Transformer (SaFT) to resolve these issues. Specifically, we equip SaFT w ith a vertical fusion module (VFM) for cross-scale semantic enhancement and a ho rizontal fusion module (HFM) for cross-sample feature fusion. Together, they bro aden the vision for each feature point from the support to a whole augmented fea ture pyramid from the query, facilitating semantic-aligned associations. Extensi ve experiments on multiple benchmarks demonstrate the superiority of our framewo rk. Without fine-tuning on novel classes, it brings significant performance gain s to one-stage baselines, lifting state-of-the-art results to a higher level.
****

Beyond Supervised vs. Unsupervised: Representative Benchmarking and Analysis of Image Representation Learning
Matthew Gwilliam, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9642-9652
By leveraging contrastive learning, clustering, and other pretext tasks, unsuper vised methods for learning image representations have reached impressive results on standard benchmarks. The result has been a crowded field -- many methods wit

h substantially different implementations yield results that seem nearly identical on popular benchmarks, such as linear evaluation on ImageNet. However, a single result does not tell the whole story. In this paper, we compare methods using performance-based benchmarks such as linear evaluation, nearest neighbor classification, and clustering for several different datasets, demonstrating the lack of a clear frontrunner within the current state-of-the-art. In contrast to prior work that performs only supervised vs. unsupervised comparison, we compare several different unsupervised methods against each other. To enrich this comparison, we analyze embeddings with measurements such as uniformity, tolerance, and centered kernel alignment (CKA), and propose two new metrics of our own: nearest neighbor graph similarity and linear prediction overlap. We reveal through our analysis that in isolation, single popular methods should not be treated as though they represent the field as a whole, and that future work ought to consider how to leverage the complimentary nature of these methods. We also leverage CKA to provide a framework to robustly quantify augmentation invariance, and provide a reminder that certain types of invariance will be undesirable for downstream tasks.

****

Few-Shot Incremental Learning for Label-to-Image Translation

Pei Chen, Yangkang Zhang, Zejian Li, Lingyun Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3697-3707

Label-to-image translation models generate images from semantic label maps. Existing models depend on large volumes of pixel-level annotated samples. When given new training samples annotated with novel semantic classes, the models should be trained from scratch with both learned and new classes. This hinders their practical applications and motivates us to introduce an incremental learning strategy to the label-to-image translation scenario. In this paper, we introduce a few-shot incremental learning method for label-to-image translation. It learns new classes one by one from a few samples of each class. We propose to adopt semantically-adaptive convolution filters and normalization. When incrementally trained on a novel semantic class, the model only learns a few extra parameters of class-specific modulation. Such design avoids catastrophic forgetting of already-learned semantic classes and enables label-to-image translation of scenes with increasingly rich content. Furthermore, to facilitate few-shot learning, we propose a modulation transfer strategy for better initialization. Extensive experiments show that our method outperforms existing related methods in most cases and achieves zero forgetting.

****

Discrete Time Convolution for Fast Event-Based Stereo

Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, Luziwei Leng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8676-8686

Inspired by biological retina, dynamical vision sensor transmits events of instantaneous changes of pixel intensity, giving it a series of advantages over traditional frame-based camera, such as high dynamical range, high temporal resolution and low power consumption. However, extracting information from highly asynchronous event data is a challenging task. Inspired by continuous dynamics of biological neuron models, we propose a novel encoding method for sparse events - continuous time convolution (CTC) - which learns to model the spatial feature of the data with intrinsic dynamics. Adopting channel-wise parameterization, temporal dynamics of the model is synchronized on the same feature map and diverges across different ones, enabling it to embed data in a variety of temporal scales. Abstracted from CTC, we further develop discrete time convolution (DTC) which accelerates the process with lower computational cost. We apply these methods to event-based multi-view stereo matching where they surpass state-of-the-art methods on benchmark criteria of the MVSEC dataset. Spatially sparse event data often leads to inaccurate estimation of edges and local contours. To address this problem, we propose a dual-path architecture in which the feature map is complemented by underlying edge information from original events extracted with spatially-adaptive denormalization. We demonstrate the superiority of our model in terms of sp

eed (up to 110 FPS), accuracy and robustness, showing a great potential for real-time fast depth estimation. Finally, we perform experiments on the recent DSEC dataset to demonstrate the general usage of our model.
****

An Image Patch Is a Wave: Phase-Aware Vision MLP
Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10935-10944
In the field of computer vision, recent works show that a pure MLP architecture mainly stacked by fully-connected layers can achieve competing performance with CNN and transformer. An input image of vision MLP is usually split into multiple tokens (patches), while the existing MLP models directly aggregate them with fixed weights, neglecting the varying semantic information of tokens from different images. To dynamically aggregate tokens, we propose to represent each token as a wave function with two parts, amplitude and phase. Amplitude is the original feature and the phase term is a complex value changing according to the semantic contents of input images. Introducing the phase term can dynamically modulate the relationship between tokens and fixed weights in MLP. Based on the wave-like token representation, we establish a novel Wave-MLP architecture for vision tasks. Extensive experiments demonstrate that the proposed Wave-MLP is superior to the state-of-the-art MLP architectures on various vision tasks such as image classification, object detection and semantic segmentation. The source code is available at https://github.com/huawei-noah/CV-Backbones/tree/master/wavemlp_pytorch and https://gitee.com/mindspore/models/tree/master/research/cv/wave_mlp.
****

Escaping Data Scarcity for High-Resolution Heterogeneous Face Hallucination
Yiqun Mei, Pengfei Guo, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18676-18686
In Heterogeneous Face Recognition (HFR), the objective is to match faces across two different domains such as visible and thermal. Large domain discrepancy makes HFR a difficult problem. Recent methods attempting to fill the gap via synthesis have achieved promising results, but their performance is still limited by the scarcity of paired training data. In practice, large-scale heterogeneous face data are often inaccessible due to the high cost of acquisition and annotation process as well as privacy regulations. In this paper, we propose a new face hallucination paradigm for HFR, which not only enables data-efficient synthesis but also allows to scale up model training without breaking any privacy policy. Unlike existing methods that learn face synthesis entirely from scratch, our approach is particularly designed to take advantage of rich and diverse facial priors from visible domain for more faithful hallucination. On the other hand, large-scale training is enabled by introducing a new federated learning scheme to allow institution-wise collaborations while avoiding explicit data sharing. Extensive experiments demonstrate the advantages of our approach in tackling HFR under current data limitations. In a unified framework, our method yields the state-of-the-art hallucination results on multiple HFR datasets.
****

Visual Acoustic Matching
Changan Chen, Ruohan Gao, Paul Calamia, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18858-18868
We introduce the visual acoustic matching task, in which an audio clip is transformed to sound like it was recorded in a target environment. Given an image of the target environment and a waveform for the source audio, the goal is to re-synthesize the audio to match the target room acoustics as suggested by its visible geometry and materials. To address this novel task, we propose a cross-modal transformer model that uses audio-visual attention to inject visual properties into the audio and generate realistic audio output. In addition, we devise a self-supervised training objective that can learn acoustic matching from in-the-wild Web videos, despite their lack of acoustically mismatched audio. We demonstrate that our approach successfully translates human speech to a variety of real-world

environments depicted in images, outperforming both traditional acoustic matching and more heavily supervised baselines.
****
Shunted Self-Attention via Multi-Scale Token Aggregation
Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10853-10862
Recent Vision Transformer (ViT) models have demonstrated encouraging results across various computer vision tasks, thanks to its competence in modeling long-range dependencies of image patches or tokens via self-attention. These models, however, usually designate the similar receptive fields of each token feature within each layer. Such a constraint inevitably limits the ability of each self-attention layer in capturing multi-scale features, thereby leading to performance degradation in handling images with multiple objects of different scales. To address this issue, we propose a novel and generic strategy, termed shunted self-attention (SSA), that allows ViTs to model the attentions at hybrid scales per attention layer. The key idea of SSA is to inject heterogeneous receptive field sizes into tokens: before computing the self-attention matrix, it selectively merges tokens to represent larger object features while keeping certain tokens to preserve fine-grained features. This novel merging scheme enables the self-attention to learn relationships between objects with different sizes and simultaneously reduces the token numbers and the computational cost. Extensive experiments across various tasks demonstrate the superiority of SSA. Specifically, the SSA-based transformer achieves 84.0% Top-1 accuracy and outperforms the state-of-the-art Focal Transformer on ImageNet with only half of the model size and computation cost, and surpasses Focal Transformer by 1.3 mAP on COCO and 2.9 mIOU on ADE20K under similar parameter and computation cost. Code has been released at \href https://github.com/OliverRensu/Shunted-Transformer https://github.com/OliverRensu/Shunted-Transformer .
****
Shadows Can Be Dangerous: Stealthy and Effective Physical-World Adversarial Attack by Natural Phenomenon
Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, Xiangyang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15345-15354
Estimating the risk level of adversarial examples is essential for safely deploying machine learning models in the real world. One popular approach for physical-world attacks is to adopt the "sticker-pasting" strategy, which however suffers from some limitations, including difficulties in access to the target or printing by valid color. A new type of non-invasive attacks emerged recently, which attempt to cast perturbation onto the target by optics based tools, such as laser beam and projector. However, the added optical patterns are artificial but not natural. Thus, they are still conspicuous and attention-grabbed, and can be easily noticed by humans. In this paper, we study a new type of optical adversarial examples, in which the perturbations are generated by a very common natural phenomenon, shadow, to achieve naturalistic and stealthy physical-world adversarial attack under the black-box setting. We extensively evaluate the effectiveness of this new attack on both simulated and real-world environments. Experimental results on traffic sign recognition demonstrate that our algorithm can generate adversarial examples effectively, reaching 98.23% and 90.47% success rates on LISA and GTSRB test sets respectively, while continuously misleading a moving camera over 95% of the time in real-world scenarios. We also offer discussions about the limitations and the defense mechanism of this attack.
****
ImplicitAtlas: Learning Deformable Shape Templates in Medical Imaging
Jiancheng Yang, Udaranga Wickramasinghe, Bingbing Ni, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15861-15871
Deep implicit shape models have become popular in the computer vision community at large but less so for biomedical applications. This is in part because large

training databases do not exist and in part because biomedical annotations are o
ften noisy. In this paper, we show that by introducing templates within the deep
 learning pipeline we can overcome these problems. The proposed framework, named
 ImplicitAtlas, represents a shape as a deformation field from a learned templat
e field, where multiple templates could be integrated to improve the shape repre
sentation capacity at negligible computational cost. Extensive experiments on th
ree medical shape datasets prove the superiority over current implicit represent
ation methods.
****

Unified Multivariate Gaussian Mixture for Efficient Neural Image Compression
Xiaosu Zhu, Jingkuan Song, Lianli Gao, Feng Zheng, Heng Tao Shen; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 17612-17621
Modeling latent variables with priors and hyperpriors is an essential problem in
 variational image compression. Formally, trade-off between rate and distortion
is handled well if priors and hyperpriors precisely describe latent variables. C
urrent practices only adopt univariate priors and process each variable individu
ally. However, we find inter-correlations and intra-correlations exist when obse
rving latent variables in a vectorized perspective. These findings reveal visual
 redundancies to improve rate-distortion performance and parallel processing abi
lity to speed up compression. This encourages us to propose a novel vectorized p
rior. Specifically, a multivariate Gaussian mixture is proposed with means and c
ovariances to be estimated. Then, a novel probabilistic vector quantization is u
tilized to effectively approximate means, and remaining covariances are further
induced to a unified mixture and solved by cascaded estimation without context m
odels involved. Furthermore, codebooks involved in quantization are extended to
multi-codebooks for complexity reduction, which formulates an efficient compress
ion procedure. Extensive experiments on benchmark datasets against state-of-the-
art indicate our model has better rate-distortion performance and an impressive
3.18xcompression speed up, giving us the ability to perform real-time, high-qual
ity variational image compression in practice. Our source code is publicly avail
able at https://github.com/xiaosu-zhu/McQuic.
****

3D Photo Stylization: Learning To Generate Stylized Novel Views From a Single Im
age
Fangzhou Mu, Jian Wang, Yicheng Wu, Yin Li; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16273-16282
Visual content creation has spurred a soaring interest given its applications in
 mobile photography and AR / VR. Style transfer and single-image 3D photography
as two representative tasks have so far evolved independently. In this paper, we
 make a connection between the two, and address the challenging task of 3D photo
 stylization - generating stylized novel views from a single image given an arbi
trary style. Our key intuition is that style transfer and view synthesis have to
 be jointly modeled. To this end, we propose a deep model that learns geometry-a
ware content features for stylization from a point cloud representation of the s
cene, resulting in high-quality stylized images that are consistent across views
. Further, we introduce a novel training protocol to enable the learning using o
nly 2D images. We demonstrate the superiority of our method via extensive qualit
ative and quantitative studies, and showcase key applications of our method in l
ight of the growing demand for 3D content creation from 2D image assets.
****

Improving Visual Grounding With Visual-Linguistic Verification and Iterative Rea
soning
Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, Weiming Hu; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp
. 9499-9508
Visual grounding is a task to locate the target indicated by a natural language
expression. Existing methods extend the generic object detection framework to th
is problem. They base the visual grounding on the features from pre-generated pr
oposals or anchors, and fuse these features with the text embeddings to locate t

he target mentioned by the text. However, modeling the visual features from thes e predefined locations may fail to fully exploit the visual context and attribut e information in the text query, which limits their performance. In this paper, we propose a transformer-based framework for accurate visual grounding by establ ishing text-conditioned discriminative features and performing multi-stage cross -modal reasoning. Specifically, we develop a visual-linguistic verification modu le to focus the visual features on regions relevant to the textual descriptions while suppressing the unrelated areas. A language-guided feature encoder is also devised to aggregate the visual contexts of the target object to improve the ob ject's distinctiveness. To retrieve the target from the encoded visual features, we further propose a multi-stage cross-modal decoder to iteratively speculate o n the correlations between the image and text for accurate target localization. Extensive experiments on five widely used datasets validate the efficacy of our proposed components and demonstrate state-of-the-art performance.

****

Contrastive Learning for Space-Time Correspondence via Self-Cycle Consistency

Jeany Son; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14679-14688

We propose a novel probabilistic method employing Bayesian Model Averaging and s elf-cycle regularization for spatio-temporal correspondence learning in videos w ithin a self-supervised learning framework. Most existing methods for self-super vised correspondence learning suffer from noisy labels that come with the data f or free, and the presence of occlusion exacerbates the problem. We tackle this i ssue within a probabilistic framework that handles model uncertainty inherent in the path selection problem built on a complete graph. We propose a self-cycle r egularization to consider a cycle-consistency property on individual edges in or der to prevent converging on noisy matching or trivial solutions. We also utiliz e a mixture of sequential Bayesian filters to estimate posterior distribution fo r targets. In addition, we present a domain contrastive loss to learn discrimina tive representation among videos. Our algorithm is evaluated on various datasets for video label propagation tasks including DAVIS2017, VIP and JHMDB, and shows outstanding performances compared to the state-of-the-art self-supervised learn ing based video correspondence algorithms. Moreover, our method converges signif icantly faster than previous methods.

****

Learning Robust Image-Based Rendering on Sparse Scene Geometry via Depth Complet ion

Yuqi Sun, Shili Zhou, Ri Cheng, Weimin Tan, Bo Yan, Lang Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7813-7823

Recent image-based rendering (IBR) methods usually adopt plenty of views to reco nstruct dense scene geometry. However, the number of available views is limited in practice. When only few views are provided, the performance of these methods drops off significantly, as the scene geometry becomes sparse as well. Therefore , in this paper, we propose Sparse-IBRNet (SIBRNet) to perform robust IBR on spa rse scene geometry by depth completion. The SIBRNet has two stages, geometry rec overy (GR) stage and light blending (LB) stage. Specifically, GR stage takes spa rse depth map and RGB as input to predict dense depth map by exploiting the corr elation between two modals. As inaccuracy of the complete depth map may cause pr ojection biases in the warping process, LB stage first uses a bias-corrected mod ule (BCM) to rectify deviations, and then aggregates modified features from diff erent views to render a novel view. Extensive experimental results demonstrate t hat our method performs best on sparse scene geometry than recent IBR methods, a nd it can generate better or comparable results as well when the geometric infor mation is dense.

****

Scale-Equivalent Distillation for Semi-Supervised Object Detection

Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 14522-14531

Recent Semi-Supervised Object Detection (SS-OD) methods are mainly based on self-training, i.e., generating hard pseudo-labels by a teacher model on unlabeled data as supervisory signals. Although they achieved certain success, the limited labeled data in semi-supervised learning scales up the challenges of object detection. We analyze the challenges these methods meet with the empirical experiment results. We find that the massive False Negative samples and inferior localization precision lack consideration. Besides, the large variance of object sizes and class imbalance (i.e., the extreme ratio between background and object) hinder the performance of prior arts. Further, we overcome these challenges by introducing a novel approach, Scale-Equivalent Distillation (SED), which is a simple yet effective end-to-end knowledge distillation framework robust to large object size variance and class imbalance. SED has several appealing benefits compared to the previous works. (1) SED imposes a consistency regularization to handle the large scale variance problem. (2) SED alleviates the noise problem from the False Negative samples and inferior localization precision. (3) A re-weighting strategy can implicitly screen the potential foreground regions of the unlabeled data to reduce the effect of class imbalance. Extensive experiments show that SED consistently outperforms the recent state-of-the-art methods on different datasets with significant margins. For example, it surpasses the supervised counterpart by more than 10 mAP when using 5% and 10% labeled data on MS-COCO.

****

Recurrent Variational Network: A Deep Learning Inverse Problem Solver Applied to the Task of Accelerated MRI Reconstruction

George Yiasemis, Jan-Jakob Sonke, Clarisa Sánchez, Jonas Teuwen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 732-741

Magnetic Resonance Imaging can produce detailed images of the anatomy and physiology of the human body that can assist doctors in diagnosing and treating pathologies such as tumours. However, MRI suffers from very long acquisition times that make it susceptible to patient motion artifacts and limit its potential to deliver dynamic treatments. Conventional approaches such as Parallel Imaging and Compressed Sensing allow for an increase in MRI acquisition speed by reconstructing MR images from sub-sampled MRI data acquired using multiple receiver coils. Recent advancements in Deep Learning combined with Parallel Imaging and Compressed Sensing techniques have the potential to produce high-fidelity reconstructions from highly accelerated MRI data. In this work we present a novel Deep Learning-based Inverse Problem solver applied to the task of Accelerated MRI Reconstruction, called the Recurrent Variational Network (RecurrentVarNet), by exploiting the properties of Convolutional Recurrent Neural Networks and unrolled algorithms for solving Inverse Problems. The RecurrentVarNet consists of multiple recurrent blocks, each responsible for one iteration of the unrolled variational optimization scheme for solving the inverse problem of multi-coil Accelerated MRI Reconstruction. Contrary to traditional approaches, the optimization steps are performed in the observation domain (k-space) instead of the image domain. Each block of the RecurrentVarNet refines the observed k-space and comprises a data consistency term and a recurrent unit which takes as input a learned hidden state and the prediction of the previous block. Our proposed method achieves new state of the art qualitative and quantitative reconstruction results on 5-fold and 10-fold accelerated data from a public multi-coil brain dataset, outperforming previous conventional and deep learning-based approaches. Our code is publicly available at https://github.com/NKI-AI/direct.

****

SelfD: Self-Learning Large-Scale Driving Policies From the Web

Jimuyang Zhang, Ruizhao Zhu, Eshed Ohn-Bar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17316-17326

Effectively utilizing the vast amounts of ego-centric navigation data that is freely available on the internet can advance generalized intelligent systems, i.e., to robustly scale across perspectives, platforms, environmental conditions, scenarios, and geographical locations. However, it is difficult to directly leverage such large amounts of unlabeled and highly diverse data for complex 3D reason

ing and planning tasks. Consequently, researchers have primarily focused on its use for various auxiliary pixel- and image-level computer vision tasks that do not consider an ultimate navigational objective. In this work, we introduce SelfD, a framework for learning scalable driving by utilizing large amounts of online monocular images. Our key idea is to leverage iterative semi-supervised training when learning imitative agents from unlabeled data. To handle unconstrained viewpoints, scenes, and camera parameters, we train an image-based model that directly learns to plan in the Bird's Eye View (BEV) space. Next, we use unlabeled data to augment the decision-making knowledge and robustness of an initially trained model via self-training. In particular, we propose a pseudo-labeling step which enables making full use of highly diverse demonstration data through "hypothetical" planning-based data augmentation. We employ a large dataset of publicly available YouTube videos to train SelfD and comprehensively analyze its generalization benefits across challenging navigation scenarios. Without requiring any additional data collection or annotation efforts, SelfD demonstrates consistent improvements (by up to 24%) in driving performance evaluation on nuScenes, Argoverse, Waymo, and CARLA.

****

"The Pedestrian Next to the Lamppost" Adaptive Object Graphs for Better Instantaneous Mapping

Avishkar Saha, Oscar Mendez, Chris Russell, Richard Bowden; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19528-19537

Estimating a semantically segmented bird's-eye-view (BEV) map from a single image has become a popular technique for autonomous control and navigation. However, they show an increase in localization error with distance from the camera. While such an increase in error is entirely expected - localization is harder at distance - much of the drop in performance can be attributed to the cues used by current texture-based models, in particular, they make heavy use of object-ground intersections (such as shadows), which become increasingly sparse and uncertain for distant objects. In this work, we address these shortcomings in BEV-mapping by learning the spatial relationship between objects in a scene. We propose a graph neural network which predicts BEV objects from a monocular image by spatially reasoning about an object within the context of other objects. Our approach sets a new state-of-the-art in BEV estimation from monocular images across three large-scale datasets, including a 50% relative improvement for objects on nuScenes.

****

Attribute Group Editing for Reliable Few-Shot Image Generation

Guanqi Ding, Xinzhe Han, Shuhui Wang, Shuzhe Wu, Xin Jin, Dandan Tu, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11194-11203

Few-shot image generation is a challenging task even using the state-of-the-art Generative Adversarial Networks (GANs). Due to the unstable GAN training process and the limited training data, the generated images are often of low quality and low diversity. In this work, we propose a new "editing-based" method, i.e., Attribute Group Editing (AGE), for few-shot image generation. The basic assumption is that any image is a collection of attributes and the editing direction for a specific attribute is shared across all categories. AGE examines the internal representation learned in GANs and identifies semantically meaningful directions. Specifically, the class embedding, i.e., the mean vector of the latent codes from a specific category, is used to represent the category-relevant attributes, and the category-irrelevant attributes are learned globally by Sparse Dictionary Learning on the difference between the sample embedding and the class embedding. Given a GAN well trained on seen categories, diverse images of unseen categories can be synthesized through editing category-irrelevant attributes while keeping category-relevant attributes unchanged. Without re-training the GAN, AGE is capable of not only producing more realistic and diverse images for downstream visual applications with limited data but achieving controllable image editing with interpretable category-irrelevant directions.

****
Surpassing the Human Accuracy: Detecting Gallbladder Cancer From USG Images With Curriculum Learning
Soumen Basu, Mayank Gupta, Pratyaksha Rana, Pankaj Gupta, Chetan Arora; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20886-20896
We explore the potential of CNN-based models for gallbladder cancer (GBC) detection from ultrasound (USG) images as no prior study is known. USG is the most common diagnostic modality for GB diseases due to its low cost and accessibility. However, USG images are challenging to analyze due to low image quality, noise, and varying viewpoints due to the handheld nature of the sensor. Our exhaustive study of state-of-the-art (SOTA) image classification techniques for the problem reveals that they often fail to learn the salient GB region due to the presence of shadows in the USG images. SOTA object detection techniques also achieve low accuracy because of spurious textures due to noise or adjacent organs. We propose GBCNet to tackle the challenges in our problem. GBCNet first extracts the regions of interest (ROIs) by detecting the GB (and not the cancer), and then uses a new multi-scale, second-order pooling architecture specializing in classifying GBC. To effectively handle spurious textures, we propose a curriculum inspired by human visual acuity, which reduces the texture biases in GBCNet. Experimental results demonstrate that GBCNet significantly outperforms SOTA CNN models, as well as the expert radiologists. Our technical innovations are generic to other USG image analysis tasks as well. Hence, as a validation, we also show the efficacy of GBCNet in detecting breast cancer from USG images. Project page with source code, trained models, and data is available at https://gbc-iitd.github.io/gbcnet.
****
CroMo: Cross-Modal Learning for Monocular Depth Estimation
Yannick Verdié, Jifei Song, Barnabé Mas, Benjamin Busam, Ales■ Leonardis, Steven McDonagh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3937-3947
Learning-based depth estimation has witnessed recent progress in multiple directions; from self-supervision using monocular video to supervised methods offering highest accuracy. Complementary to supervision, further boosts to performance and robustness are gained by combining information from multiple signals. In this paper we systematically investigate key trade-offs associated with sensor and modality design choices as well as related model training strategies. Our study leads us to a new method, capable of connecting modality-specific advantages from polarisation, Time-of-Flight and structured-light inputs. We propose a novel pipeline capable of estimating depth from monocular polarisation for which we evaluate various training signals. The inversion of differentiable analytic models thereby connects scene geometry with polarisation and ToF signals and enables self-supervised and cross-modal learning. In the absence of existing multimodal datasets, we examine our approach with a custom-made multi-modal camera rig and collect CroMo; the first dataset to consist of synchronized stereo polarisation, indirect ToF and structured-light depth, captured at video rates. Extensive experiments on challenging video scenes confirm both qualitative and quantitative pipeline advantages where we are able to outperform competitive monocular depth estimation methods.
****
Self-Supervised Object Detection From Audio-Visual Correspondence
Triantafyllos Afouras, Yuki M. Asano, Francois Fagan, Andrea Vedaldi, Florian Metze; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10575-10586
We tackle the problem of learning object detectors without supervision. Differently from weakly-supervised object detection, we do not assume image-level class labels. Instead, we extract a supervisory signal from audio-visual data, using the audio component to "teach" the object detector. While this problem is related to sound source localisation, it is considerably harder because the detector must classify the objects by type, enumerate each instance of the object, and do s

o even when the object is silent. We tackle this problem by first designing a se lf-supervised framework with a contrastive objective that jointly learns to clas sify and localise objects. Then, without using any supervision, we simply use th ese self-supervised labels and boxes to train an image-based object detector. Wi th this, we outperform previous unsupervised and weakly-supervised detectors for the task of object detection and sound source localization. We also show that w e can align this detector to ground-truth classes with as little as one label pe r pseudo-class, and show how our method can learn to detect generic objects that go beyond instruments, such as airplanes and cats.
****

Autofocus for Event Cameras
Shijie Lin, Yinqiang Zhang, Lei Yu, Bin Zhou, Xiaowei Luo, Jia Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 16344-16353
Focus control (FC) is crucial for cameras to capture sharp images in challenging real-world scenarios. The autofocus (AF) facilitates the FC by automatically ad justing the focus settings. However, due to the lack of effective AF methods for the recently introduced event cameras, their FC still relies on naive AF like m anual focus adjustments, leading to poor adaptation in challenging real-world co nditions. In particular, the inherent differences between event and frame data i n terms of sensing modality, noise, temporal resolutions, etc., bring many chall enges in designing an effective AF method for event cameras. To address these ch allenges, we develop a novel event-based autofocus framework consisting of an ev ent-specific focus measure called event rate (ER) and a robust search strategy c alled event-based golden search (EGS). To verify the performance of our method, we have collected an event-based autofocus dataset (EAD) containing well-synchro nized frames, events, and focal positions in a wide variety of challenging scene s with severe lighting and motion conditions. The experiments on this dataset an d additional real-world scenarios demonstrated the superiority of our method ove r state-of-the-art approaches in terms of efficiency and accuracy.
****

Learning Multiple Adverse Weather Removal via Two-Stage Knowledge Learning and M ulti-Contrastive Regularization: Toward a Unified Model
Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, S y-Yen Kuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17653-17662
In this paper, an ill-posed problem of multiple adverse weather removal is inves tigated. Our goal is to train a model with a 'unified' architecture and only one set of pretrained weights that can tackle multiple types of adverse weathers su ch as haze, snow, and rain simultaneously. To this end, a two-stage knowledge le arning mechanism including knowledge collation (KC) and knowledge examination (K E) based on a multi-teacher and student architecture is proposed. At the KC, the student network aims to learn the comprehensive bad weather removal problem fro m multiple well-trained teacher networks where each of them is specialized in a specific bad weather removal problem. To accomplish this process, a novel collab orative knowledge transfer is proposed. At the KE, the student model is trained without the teacher networks and examined by challenging pixel loss derived by t he ground truth. Moreover, to improve the performance of our training framework, a novel loss function called multi-contrastive knowledge regularization (MCR) l oss is proposed. Experiments on several datasets show that our student model can achieve promising performance on different bad weather removal tasks simultaneo usly.
****

Polymorphic-GAN: Generating Aligned Samples Across Multiple Domains With Learned Morph Maps
Seung Wook Kim, Karsten Kreis, Daiqing Li, Antonio Torralba, Sanja Fidler; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 10630-10640
Modern image generative models show remarkable sample quality when trained on a single domain or class of objects. In this work, we introduce a generative adver

sarial network that can simultaneously generate aligned image samples from multiple related domains. We leverage the fact that a variety of object classes share common attributes, with certain geometric differences. We propose Polymorphic-GAN which learns shared features across all domains and a per-domain morph layer to morph shared features according to each domain. In contrast to previous works, our framework allows simultaneous modelling of images with highly varying geometries, such as images of human faces, painted and artistic faces, as well as multiple different animal faces. We demonstrate that our model produces aligned samples for all domains and show how it can be used for applications such as segmentation transfer and cross-domain image editing, as well as training in low-data regimes. Additionally, we apply our Polymorphic-GAN on image-to-image translation tasks and show that we can greatly surpass previous approaches in cases where the geometric differences between domains are large.
****

Appearance and Structure Aware Robust Deep Visual Graph Matching: Attack, Defense and Beyond
Qibing Ren, Qingquan Bao, Runzhong Wang, Junchi Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15263-15272
Despite the recent breakthrough of high accuracy deep graph matching (GM) over visual images, the robustness of deep GM models is rarely studied which yet has been revealed an important issue in modern deep nets, ranging from image recognition to graph learning tasks. We first show that an adversarial attack on keypoint localities and the hidden graphs can cause significant accuracy drop to deep GM models. Accordingly, we propose our defense strategy, namely Appearance and Structure Aware Robust Graph Matching (ASAR-GM). Specifically, orthogonal to de facto adversarial training (AT), we devise the Appearance Aware Regularizer (AAR) on those appearance-similar keypoints between graphs that are likely to confuse. Experimental results show that our ASAR-GM achieves better robustness compared to AT. Moreover, our locality attack can serve as a data augmentation technique, which boosts the state-of-the-art GM models even on the clean test dataset. Code is available at https://github.com/Thinklab-SJTU/RobustMatch.
****

Super-Fibonacci Spirals: Fast, Low-Discrepancy Sampling of SO(3)
Marc Alexa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8291-8300
Super-Fibonacci spirals are an extension of Fibonacci spirals, enabling fast generation of an arbitrary but fixed number of 3D orientations. The algorithm is simple and fast. A comprehensive evaluation comparing to other methods shows that the generated sets of orientations have low discrepancy, minimal spurious components in the power spectrum, and almost identical Voronoi volumes. This makes them useful for a variety of applications in vision, robotics, machine learning, and in particular Monte Carlo sampling.
****

TrackFormer: Multi-Object Tracking With Transformers
Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, Christoph Feichtenhofer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8844-8854
The challenging task of multi-object tracking (MOT) requires simultaneous reasoning about track initialization, identity, and spatio-temporal trajectories. We formulate this task as a frame-to-frame set prediction problem and introduce TrackFormer, an end-to-end trainable MOT approach based on an encoder-decoder Transformer architecture. Our model achieves data association between frames via attention by evolving a set of track predictions through a video sequence. The Transformer decoder initializes new tracks from static object queries and autoregressively follows existing tracks in space and time with the conceptually new and identity preserving track queries. Both query types benefit from self- and encoder-decoder attention on global frame-level features, thereby omitting any additional graph optimization or modeling of motion and/or appearance. TrackFormer introduces a new tracking-by-attention paradigm and while simple in its design is able

to achieve state-of-the-art performance on the task of multi-object tracking (MOT17) and segmentation (MOTS20). The code is available at https://github.com/timmeinhardt/trackformer

****

L-Verse: Bidirectional Generation Between Image and Text

Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, Kyunghoon Bae; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16526-16536

Far beyond learning long-range interactions of natural language, transformers are becoming the de-facto standard for many vision tasks with their power and scalability. Especially with cross-modal tasks between image and text, vector quantized variational autoencoders (VQ-VAEs) are widely used to make a raw RGB image into a sequence of feature vectors. To better leverage the correlation between image and text, we propose L-Verse, a novel architecture consisting of feature-augmented variational autoencoder (AugVAE) and bidirectional auto-regressive transformer (BiART) for image-to-text and text-to-image generation. Our AugVAE shows the state-of-the-art reconstruction performance on ImageNet1K validation set, along with the robustness to unseen images in the wild. Unlike other models, BiART can distinguish between image (or text) as a conditional reference and a generation target. L-Verse can be directly used for image-to-text or text-to-image generation without any finetuning or extra object detection framework. In quantitative and qualitative experiments, L-Verse shows impressive results against previous methods in both image-to-text and text-to-image generation on MS-COCO Captions. We furthermore assess the scalability of L-Verse architecture on Conceptual Captions and present the initial result of bidirectional vision-language representation learning on general domain.

****

PanopticDepth: A Unified Framework for Depth-Aware Panoptic Segmentation

Naiyu Gao, Fei He, Jian Jia, Yanhu Shan, Haoyang Zhang, Xin Zhao, Kaiqi Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1632-1642

This paper presents a unified framework for depth-aware panoptic segmentation (DPS), which aims to reconstruct 3D scene with instance-level semantics from one single image. Prior works address this problem by simply adding a dense depth regression head to panoptic segmentation (PS) networks, resulting in two independent task branches. This neglects the mutually-beneficial relations between these two tasks, thus failing to exploit handy instance-level semantic cues to boost depth accuracy while also producing sub-optimal depth maps. To overcome these limitations, we propose a unified framework for the DPS task by applying a dynamic convolution technique to both the PS and depth prediction tasks. Specifically, instead of predicting depth for all pixels at a time, we generate instance-specific kernels to predict depth and segmentation masks for each instance. Moreover, leveraging the instance-wise depth estimation scheme, we add additional instance-level depth cues to assist with supervising the depth learning via a new depth loss. Extensive experiments on Cityscapes-DPS and SemKITTI-DPS show the effectiveness and promise of our method. We hope our unified solution to DPS can lead a new paradigm in this area.

****

3D Shape Reconstruction From 2D Images With Disentangled Attribute Flow

Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3803-3813

Reconstructing 3D shape from a single 2D image is a challenging task, which needs to estimate the detailed 3D structures based on the semantic attributes from 2D image. So far, most of the previous methods still struggle to extract semantic attributes for 3D reconstruction task. Since the semantic attributes of a single image are usually implicit and entangled with each other, it is still challenging to reconstruct 3D shape with detailed semantic structures represented by the input image. To address this problem, we propose 3DAttriFlow to disentangle and extract semantic attributes through different semantic levels in the input imag

es. These disentangled semantic attributes will be integrated into the 3D shape reconstruction process, which can provide definite guidance to the reconstructio n of specific attribute on 3D shape. As a result, the 3D decoder can explicitly capture high-level semantic features at the bottom of the network, and utilize l ow-level features at the top of the network, which allows to reconstruct more ac curate 3D shapes. Note that the explicit disentangling is learned without extra labels, where the only supervision used in our training is the input image and i ts corresponding 3D shape. Our comprehensive experiments on ShapeNet dataset dem onstrate that 3DAttriFlow outperforms the state-of-the-art shape reconstruction methods, and we also validate its generalization ability on shape completion tas k. Code is available at https://github.com/junshengzhou/3DAttriFlow.
****

Feature Statistics Mixing Regularization for Generative Adversarial Networks
Junho Kim, Yunjey Choi, Youngjung Uh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11294-11303
In generative adversarial networks, improving discriminators is one of the key c omponents for generation performance. As image classifiers are biased toward tex ture and debiasing improves accuracy, we investigate 1) if the discriminators ar e biased, and 2) if debiasing the discriminators will improve generation perform ance. Indeed, we find empirical evidence that the discriminators are sensitive t o the style (e.g., texture and color) of images. As a remedy, we propose feature statistics mixing regularization (FSMR) that encourages the discriminator's pre diction to be invariant to the styles of input images. Specifically, we generate a mixed feature of an original and a reference image in the discriminator's fea ture space and we apply regularization so that the prediction for the mixed feat ure is consistent with the prediction for the original image. We conduct extensi ve experiments to demonstrate that our regularization leads to reduced sensitivi ty to style and consistently improves the performance of various GAN architectur es on nine datasets. In addition, adding FSMR to recently proposed augmentation- based GAN methods further improves image quality. Our code is available at https ://github.com/naver-ai/FSMR.
****

Learning To Learn and Remember Super Long Multi-Domain Task Sequence
Zhenyi Wang, Li Shen, Tiehang Duan, Donglin Zhan, Le Fang, Mingchen Gao; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7982-7992
Catastrophic forgetting (CF) frequently occurs when learning with non-stationary data distribution. The CF issue remains nearly unexplored and is more challengi ng when meta-learning on a sequence of domains (datasets), called sequential dom ain meta-learning (SDML). In this work, we propose a simple yet effective learni ng to learn approach, i.e., meta optimizer, to mitigate the CF problem in SDML. We first apply the proposed meta optimizer to the simplified setting of SDML, do main-aware meta-learning, where the domain labels and boundaries are known durin g the learning process. We propose dynamically freezing the network and incorpor ating it with the proposed meta optimizer by considering the domain nature durin g meta training. In addition, we extend the meta optimizer to the more general s etting of SDML, domain-agnostic meta-learning, where domain labels and boundarie s are unknown during the learning process. We propose a domain shift detection t echnique to capture latent domain change and equip the meta optimizer with it to work in this setting. The proposed meta optimizer is versatile and can be easil y integrated with several existing meta-learning algorithms. Finally, we constru ct a challenging and large-scale benchmark consisting of 10 heterogeneous domain s with a super long task sequence consisting of 100K tasks. We perform extensive experiments on the proposed benchmark for both settings and demonstrate the eff ectiveness of our proposed method, outperforming current strong baselines by a l arge margin.
****

OpenTAL: Towards Open Set Temporal Action Localization
Wentao Bao, Qi Yu, Yu Kong; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2022, pp. 2979-2989

Temporal Action Localization (TAL) has experienced remarkable success under the supervised learning paradigm. However, existing TAL methods are rooted in the closed set assumption, which cannot handle the inevitable unknown actions in open-world scenarios. In this paper, we, for the first time, step toward the Open Set TAL (OSTAL) problem and propose a general framework OpenTAL based on Evidential Deep Learning (EDL). Specifically, the OpenTAL consists of uncertainty-aware action classification, actionness prediction, and temporal location regression. With the proposed importance-balanced EDL method, classification uncertainty is learned by collecting categorical evidence majorly from important samples. To distinguish the unknown actions from background video frames, the actionness is learned by the positive-unlabeled learning. The classification uncertainty is further calibrated by leveraging the guidance from the temporal localization quality. The OpenTAL is general to enable existing TAL models for open set scenarios, and experimental results on THUMOS14 and ActivityNet1.3 benchmarks show the effectiveness of our method. The code and pre-trained models are released at https://www.rit.edu/actionlab/opental.

****

Urban Radiance Fields

Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12932-12942

The goal of this work is to perform 3D reconstruction and novel view synthesis from data captured by scanning platforms commonly deployed for world mapping in urban outdoor environments (e.g., Street View). Given a sequence of posed RGB images and lidar sweeps acquired by cameras and scanners moving through an outdoor scene, we produce a model from which 3D surfaces can be extracted and novel RGB images can be synthesized. Our approach extends Neural Radiance Fields, which has been demonstrated to synthesize realistic novel images for small scenes in controlled settings, with new methods for leveraging asynchronously captured lidar data, for addressing exposure variation between captured images, and for leveraging predicted image segmentations to supervise densities on rays pointing at the sky. Each of these three extensions provides significant performance improvements in experiments on Street View data. Our system produces state-of-the-art 3D surface reconstructions and synthesizes higher quality novel views in comparison to both traditional methods (e.g. COLMAP) and recent neural representations (e.g. Mip-NeRF).

****

Self-Supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, Jue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18710-18719

Recent studies in deepfake detection have yielded promising results when the training and testing face forgeries are from the same dataset. However, the problem remains challenging when one tries to generalize the detector to forgeries created by unseen methods in the training dataset. This work addresses the generalizable deepfake detection from a simple principle: a generalizable representation should be sensitive to diverse types of forgeries. Following this principle, we propose to enrich the "diversity" of forgeries by synthesizing augmented forgeries with a pool of forgery configurations and strengthen the "sensitivity" to the forgeries by enforcing the model to predict the forgery configurations. To effectively explore the large forgery augmentation space, we further propose to use the adversarial training strategy to dynamically synthesize the most challenging forgeries to the current model. Through extensive experiments, we show that the proposed strategies are surprisingly effective (see Figure 1), and they could achieve superior performance than the current state-of the-art methods. Code is available at https://github.com/liangchen527/SLADD.

****

Domain-Agnostic Prior for Transfer Semantic Segmentation

Xinyue Huo, Lingxi Xie, Hengtong Hu, Wengang Zhou, Houqiang Li, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7075-7085

Unsupervised domain adaptation (UDA) is an important topic in the computer vision community. The key difficulty lies in defining a common property between the source and target domains so that the source-domain features can align with the target-domain semantics. In this paper, we present a simple and effective mechanism that regularizes cross-domain representation learning with a domain-agnostic prior (DAP) that constrains the features extracted from source and target domains to align with a domain-agnostic space. In practice, this is easily implemented as an extra loss term that requires a little extra costs. In the standard evaluation protocol of transferring synthesized data to real data, we validate the effectiveness of different types of DAP, especially one borrowed from a text embedding model that shows favorable performance beyond the state-of-the-art UDA approaches in terms of segmentation accuracy. Our research reveals that much room is left for designing better proxies for UDA.

****

Dynamic Kernel Selection for Improved Generalization and Memory Efficiency in Meta-Learning

Arnav Chavan, Rishabh Tiwari, Udbhav Bamba, Deepak K. Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9851-9860

Gradient based meta-learning methods are prone to overfit on the meta-training set, and this behaviour is more prominent with large and complex networks. Moreover, large networks restrict the application of meta-learning models on low-power edge devices. While choosing smaller networks avoid these issues to a certain extent, it affects the overall generalization leading to reduced performance. Clearly, there is an approximately optimal choice of network architecture that is best suited for every meta-learning problem, however, identifying it beforehand is not straightforward. In this paper, we present MetaDOCK, a task-specific dynamic kernel selection strategy for designing compressed CNN models that generalize well on unseen tasks in meta-learning. Our method is based on the hypothesis that for a given set of similar tasks, not all kernels of the network are needed by each individual task. Rather, each task uses only a fraction of the kernels, and the selection of the kernels per task can be learnt dynamically as a part of the inner update steps. MetaDOCK compresses the meta-model as well as the task-specific inner models, thus providing significant reduction in model size for each task, and through constraining the number of active kernels for every task, it implicitly mitigates the issue of meta-overfitting. We show that for the same inference budget, pruned versions of large CNN models obtained using our approach consistently outperform the conventional choices of CNN models. MetaDOCK couples well with popular meta-learning approaches such as iMAML. The efficacy of our method is validated on CIFAR-fs and mini-ImageNet datasets, and we have observed that our approach can provide improvements in model accuracy of up to 2% on standard meta-learning benchmark, while reducing the model size by more than 75%. Our code will be publicly available.

****

Ego4D: Around the World in 3,000 Hours of Egocentric Video

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolá■, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbel

áez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18995-19012

We introduce Ego4D, a massive-scale egocentric video dataset and benchmark suite. It offers 3,670 hours of daily-life activity video spanning hundreds of scenarios (household, outdoor, workplace, leisure, etc.) captured by 931 unique camera wearers from 74 worldwide locations and 9 different countries. The approach to collection is designed to uphold rigorous privacy and ethics standards, with consenting participants and robust de-identification procedures where relevant. Ego4D dramatically expands the volume of diverse egocentric video footage publicly available to the research community. Portions of the video are accompanied by audio, 3D meshes of the environment, eye gaze, stereo, and/or synchronized videos from multiple egocentric cameras at the same event. Furthermore, we present a host of new benchmark challenges centered around understanding the first-person visual experience in the past (querying an episodic memory), present (analyzing hand-object manipulation, audio-visual conversation, and social interactions), and future (forecasting activities). By publicly sharing this massive annotated dataset and benchmark suite, we aim to push the frontier of first-person perception. Project page: https://ego4d-data.org/

****

Differentially Private Federated Learning With Local Regularization and Sparsification

Anda Cheng, Peisong Wang, Xi Sheryl Zhang, Jian Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10122-10131

User-level differential privacy (DP) provides certifiable privacy guarantees to the information that is specific to any user's data in federated learning. Existing methods that ensure user-level DP come at the cost of severe accuracy decrease. In this paper, we study the cause of model performance degradation in federated learning with user-level DP guarantee. We find the key to solving this issue is to naturally restrict the norm of local updates before executing operations that guarantee DP. To this end, we propose two techniques, Bounded Local Update Regularization and Local Update Sparsification, to increase model quality without sacrificing privacy. We provide theoretical analysis on the convergence of our framework and give rigorous privacy guarantees. Extensive experiments show that our framework significantly improves the privacy-utility trade-off over the state-of-the-arts for federated learning with user-level DP guarantee.

****

Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, Ali Hatamizadeh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20730-20740

Vision Transformers (ViT)s have shown great performance in self-supervised learning of global and local representations that can be transferred to downstream applications. Inspired by these results, we introduce a novel self-supervised learning framework with tailored proxy tasks for medical image analysis. Specifically, we propose: (i) a new 3D transformer-based model, dubbed Swin UNEt TRansformers (Swin UNETR), with a hierarchical encoder for self-supervised pre-training; (ii) tailored proxy tasks for learning the underlying pattern of human anatomy. We demonstrate successful pre-training of the proposed model on 5050 publicly available computed tomography (CT) images from various body organs. The effectiveness of our approach is validated by fine-tuning the pre-trained models on the Beyond the Cranial Vault (BTCV) Segmentation Challenge with 13 abdominal organs and segmentation tasks from the Medical Segmentation Decathlon (MSD) dataset. Our model is currently the state-of-the-art on the public test leaderboards of both MSD and BTCV datasets. Code: https://monai.io/research/swin-unetr.

****

Camera-Conditioned Stable Feature Generation for Isolated Camera Supervised Person Re-IDentification
Chao Wu, Wenhang Ge, Ancong Wu, Xiaobin Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20238-20248
To learn camera-view invariant features for person Re-IDentification (Re-ID), the cross-camera image pairs of each person play an important role. However, such cross-view training samples could be unavailable under the ISolated Camera Supervised (ISCS) setting, e.g., a surveillance system deployed across distant scenes. To handle this challenging problem, a new pipeline is introduced by synthesizing the cross-camera samples in the feature space for model training. Specifically, the feature encoder and generator are end-to-end optimized under a novel method, Camera-Conditioned Stable Feature Generation (CCSFG). Its joint learning procedure raises concern on the stability of generative model training. Therefore, a new feature generator, Sigma-Regularized Conditional Variational Autoencoder (Sigma-Reg CVAE), is proposed with theoretical and experimental analysis on its robustness. Extensive experiments on two ISCS person Re-ID datasets demonstrate the superiority of our CCSFG to the competitors.
****

Weakly Supervised Semantic Segmentation Using Out-of-Distribution Data
Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16897-16906
Weakly supervised semantic segmentation (WSSS) methods are often built on pixel-level localization maps obtained from a classifier. However, training on class labels only, classifiers suffer from the spurious correlation between foreground and background cues (e.g. train and rail), fundamentally bounding the performance of WSSS methods. There have been previous endeavors to address this issue with additional supervision. We propose a novel source of information to distinguish foreground from the background: Out-of-Distribution (OoD) data, or images devoid of foreground object classes. In particular, we utilize the hard OoDs that the classifier is likely to make false-positive predictions. These samples typically carry key visual features on the background (e.g. rail) that the classifiers often confuse as the foreground class (e.g. train). These background cues let classifiers correctly suppress spurious background cues, resulting in an improved pixel-wise map from the classifier. From the cost point of view, acquiring such hard OoDs does not require an extensive amount of annotation efforts; it only incurs a few additional image-level labeling costs on top of the original efforts to collect the weak training set with the image labels. We propose a method, W-OoD, for utilizing the hard OoDs. W-OoD achieves state-of-the-art performance on Pascal VOC 2012. The code is available at: https://github.com/naver-ai/w-ood.
****

Point-Level Region Contrast for Object Detection Pre-Training
Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, Alexander C. Berg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16061-16070
In this work we present point-level region contrast, a self-supervised pre-training approach for the task of object detection. This approach is motivated by the two key factors in detection: localization and recognition. While accurate localization favors models that operate at the pixel- or point-level, correct recognition typically relies on a more holistic, region-level view of objects. Incorporating this perspective in pre-training, our approach performs contrastive learning by directly sampling individual point pairs from different regions. Compared to an aggregated representation per region, our approach is more robust to the change in input region quality, and further enables us to implicitly improve initial region assignments via online knowledge distillation during training. Both advantages are important when dealing with imperfect regions encountered in the unsupervised setting. Experiments show point-level region contrast improves on state-of-the-art pre-training methods for object detection and segmentation across multiple tasks and datasets, and we provide extensive ablation studies and visualizations to aid understanding. Code will be made available.

****
Upright-Net: Learning Upright Orientation for 3D Point Cloud
Xufang Pang, Feng Li, Ning Ding, Xiaopin Zhong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14911-14919
A mass of experiments shows that the pose of the input 3D models exerts a tremendous influence on automatic 3D shape analysis. In this paper, we propose Upright-Net, a deep-learning-based approach for estimating the upright orientation of 3D point clouds. Based on a well-known postulate of design states that "form ever follows function", we treat the natural base of an object as a common functional structure, which supports the object in a most commonly seen pose following a set of specific rules, e.g. physical laws, functionality-related geometric properties, semantic cues, and so on. Thus we apply a data-driven deep learning method to automatically encode those rules and formulate the upright orientation estimation problem as a classification model, i.e. extract the points on a 3D model that forms the natural base. And then the upright orientation is computed as the normal of the natural base. Our proposed new approach has three advantages. First, it formulates the continuous orientation estimation task as a discrete classification task while preserving the continuity of the solution space. Second, it automatically learns the comprehensive criteria defining a natural base of general 3D models even with asymmetric geometry. Third, the learned orientation-aware features can serve well in downstream tasks. Results show that our approach outperforms previous approaches on orientation estimation and also achieves remarkable generalization capability and transfer capability.
****
Learning Semantic Associations for Mirror Detection
Huankang Guan, Jiaying Lin, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5941-5950
Mirrors generally lack a consistent visual appearance, making mirror detection very challenging. Although recent works that are based on exploiting contextual contrasts and corresponding relations have achieved good results, heavily relying on contextual contrasts and corresponding relations to discover mirrors tend to fail in complex real-world scenes, where a lot of objects, e.g., doorways, may have similar features as mirrors. We observe that humans tend to place mirrors in relation to certain objects for specific functional purposes, e.g., a mirror above the sink. Inspired by this observation, we propose a model to exploit the semantic associations between the mirror and its surrounding objects for a reliable mirror localization. Our model first acquires class-specific knowledge of the surrounding objects via a semantic side-path. It then uses two novel modules to exploit semantic associations: 1) an Associations Exploration (AE) Module to extract the associations of the scene objects based on fully connected graph models, and 2) a Quadruple-Graph (QG) Module to facilitate the diffusion and aggregation of semantic association knowledge using graph convolutions. Extensive experiments show that our method outperforms the existing methods and sets the new state-of-the-art on both PMD dataset (f-measure: 0.844) and MSD dataset (f-measure: 0.889).
****
Spatial-Temporal Parallel Transformer for Arm-Hand Dynamic Estimation
Shuying Liu, Wenbin Wu, Jiaxian Wu, Yue Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20523-20532
We propose an approach to estimate arm and hand dynamics from monocular video by utilizing the relationship between arm and hand. Although monocular full human motion capture technologies have made great progress in recent years, recovering accurate and plausible arm twists and hand gestures from in-the-wild videos still remains a challenge. To solve this problem, our solution is proposed based on the fact that arm poses and hand gestures are highly correlated in most real situations. To make best use of arm-hand correlation as well as inter-frame information, we carefully design a Spatial-Temporal Parallel Arm-Hand Motion Transformer (PAHMT) to predict the arm and hand dynamics simultaneously. We also introduce new losses to encourage the estimations to be smooth and accurate. Besides, we collect a motion capture dataset including 200K frames of hand gestures and use

this data to train our model. By integrating a 2D hand pose estimation model and a 3D human pose estimation model, the proposed method can produce plausible arm and hand dynamics from monocular video. Extensive evaluations demonstrate that the proposed method has advantages over previous state-of-the-art approaches and shows robustness under various challenging scenarios.

****

Failure Modes of Domain Generalization Algorithms

Tigran Galstyan, Hrayr Harutyunyan, Hrant Khachatrian, Greg Ver Steeg, Aram Galstyan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19077-19086

Domain generalization algorithms use training data from multiple domains to learn models that generalize well to unseen domains. While recently proposed benchmarks demonstrate that most of the existing algorithms do not outperform simple baselines, the established evaluation methods fail to expose the impact of various factors that contribute to the poor performance. In this paper we propose an evaluation framework for domain generalization algorithms that allows decomposition of the error into components capturing distinct aspects of generalization. Inspired by the prevalence of algorithms based on the idea of domain-invariant representation learning, we extend the evaluation framework to capture various types of failures in achieving invariance. We show that the largest contributor to the generalization error varies across methods, datasets, regularization strengths and even training lengths. We observe two problems associated with the strategy of learning domain-invariant representations. On Colored MNIST, most domain generalization algorithms fail because they reach domain-invariance only on the training domains. On Camelyon-17, domain-invariance degrades the quality of representations on unseen domains. We hypothesize that focusing instead on tuning the classifier on top of a rich representation can be a promising direction.

****

Geometric and Textural Augmentation for Domain Gap Reduction

Xiao-Chang Liu, Yong-Liang Yang, Peter Hall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14340-14350

Research has shown that convolutional neural networks for object recognition are vulnerable to changes in depiction because learning is biased towards the low-level statistics of texture patches. Recent works concentrate on improving robustness by applying style transfer to training examples to mitigate against over-fitting to one depiction style. These new approaches improve performance, but they ignore the geometric variations in object shape that real art exhibits: artists deform and warp objects for artistic effect. Motivated by this observation, we propose a method to reduce bias by jointly increasing the texture and geometry diversities of the training data. In effect, we extend the visual object class to include examples with shape changes that artists use. Specifically, we learn the distribution of warps that cover each given object class. Together with augmenting textures based on a broad distribution of styles, we show by experiments that our method improves performance on several cross-domain benchmarks.

****

Class Similarity Weighted Knowledge Distillation for Continual Semantic Segmentation

Minh Hieu Phan, The-Anh Ta, Son Lam Phung, Long Tran-Thanh, Abdesselam Bouzerdoum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16866-16875

Deep learning models are known to suffer from the problem of catastrophic forgetting when they incrementally learn new classes. Continual learning for semantic segmentation (CSS) is an emerging field in computer vision. We identify a problem in CSS: A model tends to be confused between old and new classes that are visually similar, which makes it forget the old ones. To address this gap, we propose REMINDER - a new CSS framework and a novel class similarity knowledge distillation (CSW-KD) method. Our CSW-KD method distills the knowledge of a previous model on old classes that are similar to the new one. This provides two main benefits: (i) selectively revising old classes that are more likely to be forgotten, and (ii) better learning new classes by relating them with the previously seen cl

asses. Extensive experiments on Pascal-VOC 2012 and ADE20k datasets show that ou r approach outperforms state-of-the-art methods on standard CSS settings by up t o 7.07% and 8.49%, respectively.
****

DAD-3DHeads: A Large-Scale Dense, Accurate and Diverse Dataset for 3D Head Align ment From a Single Image
Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Ji■í Matas, Viktor iia Sharmanska; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2022, pp. 20942-20952
We present DAD-3DHeads, a dense and diverse large-scale dataset, and a robust mo del for 3D Dense Head Alignment in-the-wild. It contains annotations of over 3.5 K landmarks that accurately represent 3D head shape compared to the ground-truth scans. The data-driven model, DAD-3DNet, trained on our dataset, learns shape, expression, and pose parameters, and performs 3D reconstruction of a FLAME mesh. The model also incorporates a landmark prediction branch to take advantage of r ich supervision and co-training of multiple related tasks. Experimentally, DAD- 3DNet outperforms or is comparable to the state-of-the-art models in (i) 3D Head Pose Estimation on AFLW2000-3D and BIWI, (ii) 3D Face Shape Reconstruction on N oW and Feng, and (iii) 3D Dense Head Alignment and 3D Landmarks Estimation on DA D-3DHeads dataset. Finally, diversity of DAD-3DHeads in camera angles, facial ex pressions, and occlusions enables a benchmark to study in-the-wild generalizatio n and robustness to distribution shifts. The dataset webpage is https://p.farm/r esearch/dad-3dheads.
****

Reconstructing Surfaces for Sparse Point Clouds With On-Surface Priors
Baorui Ma, Yu-Shen Liu, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6315-6325
It is an important task to reconstruct surfaces from 3D point clouds. Current me thods are able to reconstruct surfaces by learning Signed Distance Functions (SD Fs) from single point clouds without ground truth signed distances or point norm als. However, they require the point clouds to be dense, which dramatically limi ts their performance in real applications. To resolve this issue, we propose to reconstruct highly accurate surfaces from sparse point clouds with an on-surface prior. We train a neural network to learn SDFs via projecting queries onto the surface represented by the sparse point cloud. Our key idea is to infer signed d istances by pushing both the query projections to be on the surface and the proj ection distance to be the minimum. To achieve this, we train a neural network to capture the on-surface prior to determine whether a point is on a sparse point cloud or not, and then leverage it as a differentiable function to learn SDFs fr om unseen sparse point cloud. Our method can learn SDFs from a single sparse poi nt cloud without ground truth signed distances or point normals. Our numerical e valuation under widely used benchmarks demonstrates that our method achieves sta te-of-the-art reconstruction accuracy, especially for sparse point clouds. Code and data are available at https://github.com/mabaorui/OnSurfacePrior.
****

HybridCR: Weakly-Supervised 3D Point Cloud Semantic Segmentation via Hybrid Cont rastive Regularization
Mengtian Li, Yuan Xie, Yunhang Shen, Bo Ke, Ruizhi Qiao, Bo Ren, Shaohui Lin, Li zhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14930-14939
To address the huge labeling cost in large-scale point cloud semantic segmentati on, we propose a novel hybrid contrastive regularization (HybridCR) framework in weakly-supervised setting, which obtains competitive performance compared to it s fully-supervised counterpart. Specifically, HybridCR is the first framework to leverage both point consistency and employ contrastive regularization with pseu do labeling in an end-to-end manner. Fundamentally, HybridCR explicitly and effe ctively considers the semantic similarity between local neighboring points and g lobal characteristics of 3D classes. We further design a dynamic point cloud aug mentor to generate diversity and robust sample views, whose transformation param eter is jointly optimized with model training. Through extensive experiments, Hy

bridCR achieves significant performance improvement against the SOTA methods on both indoor and outdoor datasets, e.g., S3DIS, ScanNet-V2, Semantic3D, and SemanticKITTI.
****

Fine-Tuning Image Transformers Using Learnable Memory
Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, Andrew Jackson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12155-12164
In this paper we propose augmenting Vision Transformer models with learnable memory tokens. Our approach allows the model to adapt to new tasks, using few parameters, while optionally preserving its capabilities on previously learned tasks. At each layer we introduce a set of learnable embedding vectors that provide contextual information useful for specific datasets. We call these 'memory tokens'. We show that augmenting a model with just a handful of such tokens per layer significantly improves accuracy when compared to conventional head-only fine-tuning, and performs only slightly below the significantly more expensive full fine-tuning. We then propose an attention-masking approach that enables models to preserve their previous capabilities, while extending them to new downstream tasks. This approach, which we call 'non-destructive fine-tuning', enables computation reuse across multiple tasks while being able to learn new tasks independently.
****

Contrastive Conditional Neural Processes
Zesheng Ye, Lina Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9687-9696
Conditional Neural Processes(CNPs) bridge neural networks with probabilistic inference to approximate functions of Stochastic Processes under meta-learning settings. Given a batch of non-i.i.d function instantiations, CNPs are jointly optimized for in-instantiation observation prediction and cross-instantiation meta-representation adaptation within a generative reconstruction pipeline. There can be a challenge in tying together such two targets when the distribution of function observations scales to high-dimensional and noisy spaces. Instead, noise contrastive estimation might be able to provide more robust representations by learning distributional matching objectives to combat such inherent limitation of generative models. In light of this, we propose to equip CNPs by 1) aligning prediction with encoded ground-truth observation, and 2) decoupling meta-representation adaptation from generative reconstruction. Specifically, two auxiliary contrastive branches are set up hierarchically, namely in-instantiation temporal contrastive learning (TCL) and cross-instantiation function contrastive learning (FCL), to facilitate local predictive alignment and global function consistency, respectively. We empirically show that TCL captures high-level abstraction of observations, whereas FCL helps identify underlying functions, which in turn provides more efficient representations. Our model outperforms other CNPs variants when evaluating function distribution reconstruction and parameter identification across 1D, 2D and high-dimensional time-series.
****

vCLIMB: A Novel Video Class Incremental Learning Benchmark
Andrés Villa, Kumail Alhamoud, Victor Escorcia, Fabian Caba, Juan León Alcázar, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19035-19044
Continual learning (CL) is under-explored in the video domain. The few existing works contain splits with imbalanced class distributions over the tasks, or study the problem in unsuitable datasets. We introduce vCLIMB, a novel video continual learning benchmark. vCLIMB is a standardized test-bed to analyze catastrophic forgetting of deep models in video continual learning. In contrast to previous work, we focus on class incremental continual learning with models trained on a sequence of disjoint tasks, and distribute the number of classes uniformly across the tasks. We perform in-depth evaluations of existing CL methods in vCLIMB, and observe two unique challenges in video data. The selection of instances to store in episodic memory is performed at the frame level. Second, untrimmed training data influences the effectiveness of frame sampling strategies. We address th

ese two challenges by proposing a temporal consistency regularization that can b
e applied on top of memory-based continual learning methods. Our approach signif
icantly improves the baseline, by up to 24% on the untrimmed continual learning
task. The code of our benchmark can be found at: https://vclimb.netlify.app/.
****

Bending Reality: Distortion-Aware Transformers for Adapting to Panoramic Semanti
c Segmentation
Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, Rainer Stiefel
hagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2022, pp. 16917-16927
Panoramic images with their 360deg directional view encompass exhaustive informa
tion about the surrounding space, providing a rich foundation for scene understa
nding. To unfold this potential in the form of robust panoramic segmentation mod
els, large quantities of expensive, pixel-wise annotations are crucial for succe
ss. Such annotations are available, but predominantly for narrow-angle, pinhole-
camera images which, off the shelf, serve as sub-optimal resources for training
panoramic models. Distortions and the distinct image-feature distribution in 360
deg panoramas impede the transfer from the annotation-rich pinhole domain and th
erefore come with a big dent in performance. To get around this domain differenc
e and bring together semantic annotations from pinhole- and 360deg surround-visu
als, we propose to learn object deformations and panoramic image distortions in
the Deformable Patch Embedding (DPE) and Deformable MLP (DMLP) components which
blend into our Transformer for PAnoramic Semantic Segmentation (Trans4PASS) mode
l. Finally, we tie together shared semantics in pinhole- and panoramic feature e
mbeddings by generating multi-scale prototype features and aligning them in our
Mutual Prototypical Adaptation (MPA) for unsupervised domain adaptation. On the
indoor Stanford2D3D dataset, our Trans4PASS with MPA maintains comparable perfor
mance to fully-supervised state-of-the-arts, cutting the need for over 1,400 lab
eled panoramas. On the outdoor DensePASS dataset, we break state-of-the-art by 1
4.39% mIoU and set the new bar at 56.38%.
****

Sparse and Complete Latent Organization for Geospatial Semantic Segmentation
Fengyu Yang, Chenyang Ma; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2022, pp. 1809-1818
Geospatial semantic segmentation on remote sensing images suffers from large int
ra-class variance in both foreground and background classes. First, foreground o
bjects are tiny in the remote sensing images and are represented by only a few p
ixels, which leads to large foreground intra-class variance and undermines the d
iscrimination between foreground classes (issue firstly considered in this work)
. Second, background class contains complex context, which results in false alar
ms due to large background intra-class variance. To alleviate these two issues,
we construct a sparse and complete latent structure via prototypes. In particula
r, to enhance the sparsity of the latent space, we design a prototypical contras
tive learning to have prototypes of the same category clustering together and pr
ototypes of different categories to be far away from each other. Also, we streng
then the completeness of the latent space by modeling all foreground categories
and hardest (nearest) background objects. We further design a patch shuffle augm
entation for remote sensing images with complicated contexts. Our augmentation e
ncourages the semantic information of an object to be correlated only to the lim
ited context within the patch that is specific to its category, which further re
duces large intra-class variance. We conduct extensive evaluations on a large sc
ale remote sensing dataset, showing our approach significantly outperforms state
-of-the-art methods by a large margin.
****

Robust Equivariant Imaging: A Fully Unsupervised Framework for Learning To Image
 From Noisy and Partial Measurements
Dongdong Chen, Julián Tachella, Mike E. Davies; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5647-5656
Deep networks provide state-of-the-art performance in multiple imaging inverse p
roblems ranging from medical imaging to computational photography. However, most

existing networks are trained with clean signals which are often hard or impossible to obtain. Equivariant imaging (EI) is a recent self-supervised learning framework that exploits the group invariance present in signal distributions to learn a reconstruction function from partial measurement data alone. While EI results are impressive, its performance degrades with increasing noise. In this paper, we propose a Robust Equivariant Imaging (REI) framework which can learn to image from noisy partial measurements alone. The proposed method uses Stein's Unbiased Risk Estimator (SURE) to obtain a fully unsupervised training loss that is robust to noise. We show that REI leads to considerable performance gains on linear and nonlinear inverse problems, thereby paving the way for robust unsupervised imaging with deep networks. Code is available at https://github.com/edongdongchen/REI.

****

Not All Relations Are Equal: Mining Informative Labels for Scene Graph Generation

Arushi Goel, Basura Fernando, Frank Keller, Hakan Bilen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15596-15606

Scene graph generation (SGG) aims to capture a wide variety of interactions between pairs of objects, which is essential for full scene understanding. Existing SGG methods trained on the entire set of relations fail to acquire complex reasoning about visual and textual correlations due to various biases in training data. Learning on trivial relations that indicate generic spatial configuration like 'on' instead of informative relations such as 'parked on' does not enforce this complex reasoning, harming generalization. To address this problem, we propose a novel framework for SGG training that exploits relation labels based on their informativeness. Our model-agnostic training procedure imputes missing informative relations for less informative samples in the training data and trains a SGG model on the imputed labels along with existing annotations. We show that this approach can successfully be used in conjunction with state-of-the-art SGG methods and improves their performance significantly in multiple metrics on the standard Visual Genome benchmark. Furthermore, we obtain considerable improvements for unseen triplets in a more challenging zero-shot setting.

****

Learning To Detect Scene Landmarks for Camera Localization

Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, Sudipta N. Sinha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11132-11142

Modern camera localization methods that use image retrieval, feature matching, and 3D structure-based pose estimation require long-term storage of numerous scene images or a vast amount of image features. This can make them unsuitable for resource constrained VR/AR devices and also raises serious privacy concerns. We present a new learned camera localization technique that eliminates the need to store features or a detailed 3D point cloud. Our key idea is to implicitly encode the appearance of a sparse yet salient set of 3D scene points into a convolutional neural network (CNN) that can detect these scene points in query images whenever they are visible. We refer to these points as scene landmarks. We also show that a CNN can be trained to regress bearing vectors for such landmarks even when they are not within the camera's field-of-view. We demonstrate that the predicted landmarks yield accurate pose estimates and that our method outperforms DSAC*, the state-of-the-art in learned localization. Furthermore, extending HLoc (an accurate method) by combining its correspondences with our predictions, boosts its accuracy even further.

****

INS-Conv: Incremental Sparse Convolution for Online 3D Segmentation

Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18975-18984

We propose INS-Conv, an INcremental Sparse Convolutional network which enables online accurate 3D semantic and instance segmentation. Benefiting from the increm

ental nature of RGB-D reconstruction, we only need to update the residuals betwe
en the reconstructed scenes of consecutive frames, which are usually sparse. For
 layer design, we define novel residual propagation rules for sparse convolution
 operations, achieving close approximation to standard sparse convolution. For n
etwork architecture, an uncertainty term is proposed to adaptively select which
 residual to update, further improving the inference accuracy and efficiency. Bas
ed on INS-Conv, an online joint 3D semantic and instance segmentation pipeline i
s proposed, reaching an inference speed of 15 FPS on GPU and 10 FPS on CPU. Expe
riments on ScanNetv2 and SceneNN datasets show that the accuracy of our method s
urpasses previous online methods by a large margin, and is on par with state-of-
the-art offline methods. A live demo on portable devices further shows the super
ior performance of INS-Conv.
****

ST++: Make Self-Training Work Better for Semi-Supervised Semantic Segmentation
Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4268-42
77
Self-training via pseudo labeling is a conventional, simple, and popular pipelin
e to leverage unlabeled data. In this work, we first construct a strong baseline
 of self-training (namely ST) for semi-supervised semantic segmentation via inje
cting strong data augmentations (SDA) on unlabeled images to alleviate overfitti
ng noisy labels as well as decouple similar predictions between the teacher and
student. With this simple mechanism, our ST outperforms all existing methods wit
hout any bells and whistles, e.g., iterative re-training. Inspired by the impres
sive results, we thoroughly investigate the SDA and provide some empirical analy
sis. Nevertheless, incorrect pseudo labels are still prone to accumulate and deg
rade the performance. To this end, we further propose an advanced self-training
framework (namely ST++), that performs selective re-training via prioritizing re
liable unlabeled images based on holistic prediction-level stability. Concretely
, several model checkpoints are saved in the first stage supervised training, an
d the discrepancy of their predictions on the unlabeled image serves as a measur
ement for reliability. Our image-level selection offers holistic contextual info
rmation for learning. We demonstrate that it is more suitable for segmentation t
han common pixel-wise selection. As a result, ST++ further boosts the performanc
e of our ST. Code is available at https://github.com/LiheYoung/ST-PlusPlus.
****

Visual Vibration Tomography: Estimating Interior Material Properties From Monocu
lar Video
Berthy T. Feng, Alexander C. Ogren, Chiara Daraio, Katherine L. Bouman; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2022, pp. 16231-16240
An object's interior material properties, while invisible to the human eye, dete
rmine motion observed on its surface. We propose an approach that estimates hete
rogeneous material properties of an object from a monocular video of its surface
 vibrations. Specifically, we show how to estimate Young's modulus and density t
hroughout a 3D object with known geometry. Knowledge of how these values change
across the object is useful for simulating its motion and characterizing any def
ects. Traditional non-destructive testing approaches, which often require expens
ive instruments, generally estimate only homogenized material properties or simp
ly identify the presence of defects. In contrast, our approach leverages monocul
ar video to (1) identify image-space modes from an object's sub-pixel motion, an
d (2) directly infer spatially-varying Young's modulus and density values from t
he observed modes. We demonstrate our approach on both simulated and real videos
.
****

Self-Supervised Global-Local Structure Modeling for Point Cloud Domain Adaptatio
n With Reliable Voted Pseudo Labels
Hehe Fan, Xiaojun Chang, Wanyue Zhang, Yi Cheng, Ying Sun, Mohan Kankanhalli; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2022, pp. 6377-6386

In this paper, we propose an unsupervised domain adaptation method for deep poin t cloud representation learning. To model the internal structures in target poin t clouds, we first propose to learn the global representations of unlabeled data by scaling up or down point clouds and then predicting the scales. Second, to c apture the local structure in a self-supervised manner, we propose to project a 3D local area onto a 2D plane and then learn to reconstruct the squeezed region. Moreover, to effectively transfer the knowledge from source domain, we propose to vote pseudo labels for target samples based on the labels of their nearest so urce neighbors in the shared feature space. To avoid the noise caused by incorre ct pseudo labels, we only select reliable target samples, whose voting consisten cies are high enough, for enhancing adaptation. The voting method is able to ada ptively select more and more target samples during training, which in return fac ilitates adaptation because the amount of labeled target data increases. Experim ents on PointDA (ModelNet-10, ShapeNet-10 and ScanNet-10) and Sim-to-Real (Model Net-11, ScanObjectNN-11, ShapeNet-9 and ScanObjectNN-9) demonstrate the effectiv eness of our method.

****

Interacting Attention Graph for Single Image Two-Hand Reconstruction

Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, Yebin Liu ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 2761-2770

Graph convolutional network (GCN) has achieved great success in single hand reco nstruction task, while interacting two-hand reconstruction by GCN remains unexpl ored. In this paper, we present Interacting Attention Graph Hand (IntagHand), th e first graph convolution based network that reconstructs two interacting hands from a single RGB image. To solve occlusion and interaction challenges of two-ha nd reconstruction, we introduce two novel attention based modules in each upsamp ling step of the original GCN. The first module is the pyramid image feature att ention (PIFA) module, which utilizes multiresolution features to implicitly obta in vertex-to-image alignment. The second module is the cross hand attention (CHA) module that encodes the coherence of interacting hands by building dense cross -attention between two hand vertices. As a result, our model outperforms all exi sting two-hand reconstruction methods by a large margin on InterHand2.6M benchma rk. Moreover, ablation studies verify the effectiveness of both PIFA and CHA mod ules for improving the reconstruction accuracy. Results on in-the-wild images an d live video streams further demonstrate the generalization ability of our netwo rk. Our code is available at https://github.com/Dw1010/IntagHand.

****

Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task

Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan , Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2022, pp. 21341-21350

Concurrent perception datasets for autonomous driving are mainly limited to fron tal view with sensors mounted on the vehicle. None of them is designed for the o verlooked roadside perception tasks. On the other hand, the data captured from r oadside cameras have strengths over frontal-view data, which is believed to faci litate a safer and more intelligent autonomous driving system. To accelerate the progress of roadside perception, we present the first high-diversity challengin g Roadside Perception 3D dataset- Rope3D from a novel view. The dataset consists of 50k images and over 1.5M 3D objects in various scenes, which are captured un der different settings including various cameras with ambiguous mounting positio ns, camera specifications, viewpoints, and different environmental conditions. W e conduct strict 2D-3D joint annotation and comprehensive data analysis, as well as set up a new 3D roadside perception benchmark with metrics and evaluation de vkit. Furthermore, we tailor the existing frontal-view monocular 3D object detec tion approaches and propose to leverage the geometry constraint to solve the inh erent ambiguities caused by various sensors, viewpoints. Our dataset is availabl e on https://thudair.baai.ac.cn/rope.

****

Noisy Boundaries: Lemon or Lemonade for Semi-Supervised Instance Segmentation?
Zhenyu Wang, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16826-16835
Current instance segmentation methods rely heavily on pixel-level annotated images. The huge cost to obtain such fully-annotated images restricts the dataset scale and limits the performance. In this paper, we formally address semi-supervised instance segmentation, where unlabeled images are employed to boost the performance. We construct a framework for semi-supervised instance segmentation by assigning pixel-level pseudo labels. Under this framework, we point out that noisy boundaries associated with pseudo labels are double-edged. We propose to exploit and resist them in a unified manner simultaneously: 1) To combat the negative effects of noisy boundaries, we propose a noise-tolerant mask head by leveraging low-resolution features. 2) To enhance the positive impacts, we introduce a boundary-preserving map for learning detailed information within boundary-relevant regions. We evaluate our approach by extensive experiments. It behaves extraordinarily, outperforming the supervised baseline by a large margin, more than 6% on Cityscapes, 7% on COCO and 4.5% on BDD100k. On Cityscapes, our method achieves comparable performance by utilizing only 30% labeled images.
****

Boosting View Synthesis With Residual Transfer
Xuejian Rong, Jia-Bin Huang, Ayush Saraf, Changil Kim, Johannes Kopf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19760-19769
Volumetric view synthesis methods with neural representations, such as NeRF and NeX, have recently demonstrated high-quality novel view synthesis. Optimizing these representations is slow, however, and even fully trained models cannot reproduce all fine details in the input views. We present a simple but effective technique to boost the rendering quality, which can be easily integrated with most view synthesis methods. The core idea is to adaptively transfer color residuals (the difference between the input images and their reconstruction) from training views to novel views. We blend the residuals from multiple views using a heuristic weighting scheme depending on ray visibility and angular differences. We integrate our technique with several state-of-the-art view synthesis methods and evaluate on the Real Forward-facing and the Shiny datasets. Our results show that at about 1/10th the number of training iterations we achieve the same rendering quality as fully converged NeRF and NeX models, and when applied to fully converged models we significantly improve their rendering quality.
****

Input-Level Inductive Biases for 3D Reconstruction
Wang Yifan, Carl Doersch, Relja Arandjelovi■, João Carreira, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6176-6186
Much of the recent progress in 3D vision has been driven by the development of specialized architectures that incorporate geometrical inductive biases. In this paper we tackle 3D reconstruction using a domain agnostic architecture and study how instead to inject the same type of inductive biases directly as extra inputs to the model. This approach makes it possible to apply existing general models, such as Perceivers, on this rich domain, without the need for architectural changes, while simultaneously maintaining data efficiency of bespoke models. In particular we study how to encode cameras, projective ray incidence and epipolar geometry as model inputs, and demonstrate competitive multi-view depth estimation performance on multiple benchmarks.
****

Exploring and Evaluating Image Restoration Potential in Dynamic Scenes
Cheng Zhang, Shaolin Su, Yu Zhu, Qingsen Yan, Jinqiu Sun, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2067-2076
In dynamic scenes, images often suffer from dynamic blur due to superposition of motions or low signal-noise ratio resulted from quick shutter speed when avoiding motions. Recovering sharp and clean result from the captured images heavily d

epends on the ability of restoration methods and the quality of the input. Thoug h existing research on image restoration focuses on developing models for obtain ing better restored results, less have studied to evaluate how and which input i mage leads to superior restored quality. In this paper, to better study an image 's potential value that can be explored for restoration, we propose a novel conc ept, referring to image restoration potential (IRP). Specifically, We first esta blish a dynamic scene imaging dataset containing composite distortions and appli ed image restoration processes to validate the rationality of the existence to I RP. Based on this dataset, we investigate into several properties of IRP and pro pose a novel deep model to accurately predict IRP values. By gradually distillin g and selective fusing the degradation features, the proposed model shows its su periority in IRP prediction. Thanks to the proposed model, we are then able to v alidate how various image restoration related applications are benefited from IR P prediction. We show the potential usages of IRP as a filtering principle to se lect valuable frames, an auxiliary guidance to improve restoration models, and a lso an indicator to optimize camera settings for capturing better images under d ynamic scenarios.

****

FashionVLP: Vision Language Transformer for Fashion Retrieval With Feedback
Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, Pradeep Natarajan; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 14105-14115

Fashion image retrieval based on a query pair of reference image and natural lan guage feedback is a challenging task that requires models to assess fashion rela ted information from visual and textual modalities simultaneously. We propose a new vision-language transformer based model, FashionVLP, that brings the prior k nowledge contained in large image-text corpora to the domain of fashion image re -trieval, and combines visual information from multiple levels of context to eff ectively capture fashion related information. While queries are encoded through the transformer layers, our asymmetric design adopts a novel attention-based app roach for fusing target image features without involving text or transformer lay ers in the process. Extensive results show that FashionVLP achieves the state-of -the-art performance on benchmark datasets, with a large 23% relative improvemen t on the challenging FashionIQ dataset, which contains complex natural language feedback.

****

Cross-Image Relational Knowledge Distillation for Semantic Segmentation
Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, Qian Zhang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 12319-12328

Current Knowledge Distillation (KD) methods for semantic segmentation often guid e the student to mimic the teacher's structured information generated from indiv idual data samples. However, they ignore the global semantic relations among pix els across various images that are valuable for KD. This paper proposes a novel Cross-Image Relational KD (CIRKD), which focuses on transferring structured pixe l-to-pixel and pixel-to-region relations among the whole images. The motivation is that a good teacher network could construct a well-structured feature space i n terms of global pixel dependencies. CIRKD makes the student mimic better struc tured semantic relations from the teacher, thus improving the segmentation perfo rmance. Experimental results over Cityscapes, CamVid and Pascal VOC datasets dem onstrate the effectiveness of our proposed approach against state-of-the-art dis tillation methods. The code is available at https://github.com/winycg/CIRKD.

****

A-ViT: Adaptive Tokens for Efficient Vision Transformer
Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, Pavlo Molchan ov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2022, pp. 10809-10818

We introduce A-ViT, a method that adaptively adjusts the inference cost of visio n transformer ViT for images of different complexity. A-ViT achieves this by aut omatically reducing the number of tokens in vision transformers that are process

ed in the network as inference proceeds. We reformulate Adaptive Computation Time (ACT) for this task, extending halting to discard redundant spatial tokens. The appealing architectural properties of vision transformers enables our adaptive token reduction mechanism to speed up inference without modifying the network architecture or inference hardware. We demonstrate that A-ViT requires no extra parameters or sub-network for halting, as we base the learning of adaptive halting on the original network parameters. We further introduce distributional prior regularization that stabilizes training compared to prior ACT approaches. On the image classification task (ImageNet1K), we show that our proposed A-ViT yields high efficacy in filtering informative spatial features and cutting down on the overall compute. The proposed method improves the throughput of DeiT-Tiny by 62% and DeiT-Small by 38% with only 0.3% accuracy drop, outperforming prior art by a large margin.

****

Think Global, Act Local: Dual-Scale Graph Transformer for Vision-and-Language Navigation

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, Ivan Laptev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16537-16547

Following language instructions to navigate in unseen environments is a challenging problem for autonomous embodied agents. The agent not only needs to ground languages in visual scenes, but also should explore the environment to reach its target. In this work, we propose a dual-scale graph transformer (DUET) for joint long-term action planning and fine-grained cross-modal understanding. We build a topological map on-the-fly to enable efficient exploration in global action space. To balance the complexity of large action space reasoning and fine-grained language grounding, we dynamically combine a fine-scale encoding over local observations and a coarse-scale encoding on a global map via graph transformers. The proposed approach, DUET, significantly outperforms state-of-the-art methods on goal-oriented vision-and-language navigation (VLN) benchmarks REVERIE and SOON. It also improves the success rate on the fine-grained VLN benchmark R2R.

****

Towards Layer-Wise Image Vectorization

Xu Ma, Yuqian Zhou, Xingqian Xu, Bin Sun, Valerii Filev, Nikita Orlov, Yun Fu, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16314-16323

Image rasterization is a mature technique in computer graphics, while image vectorization, the reverse path of rasterization, remains a major challenge. Recent ad- vanced deep learning-based models achieve vectorization and semantic interpolation of vector graphs and demon- strate a better topology of generating new figures. How- ever, deep models cannot be easily generalized to out-of- domain testing data. The generated SVGs also contain complex and redundant shapes that are not quite conve- nient for further editing. Specifically, the crucial layer- wise topology and fundamental semantics in images are still not well understood and thus not fully explored. In this work, we propose Layer-wise Image Vectorization, namely LIVE, to convert raster images to SVGs and simultaneously maintain its image topology. LIVE can generate compact SVG forms with layer-wise structures that are semantically consistent with the human perspective. We progressively add new bezier paths and optimize these paths with the layer-wise framework, newly designed loss functions, and component-wise path initialization technique. Our experi- ments demonstrate that LIVE presents more plausible vec- torized forms than prior works and can be generalized to new images. With the help of this newly learned topol- ogy, LIVE initiates human editable SVGs for both design- ers and other downstream applications. Codes are made available at https://github.com/Picsart-AI-Research/LIVE- Layerwise-Image-Vectorization.

****

Scenic: A JAX Library for Computer Vision Research and Beyond

Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, Yi Tay; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21393-21398

Scenic is an open-source (https://github.com/google-research/scenic) JAX library with a focus on transformer-based models for computer vision research and beyond. The goal of this toolkit is to facilitate rapid experimentation, prototyping, and research of new architectures and models. Scenic supports a diverse range of tasks (e.g., classification, segmentation, detection) and facilitates working on multi-modal problems, along with GPU/TPU support for large-scale, multi-host and multi-device training. Scenic also offers optimized implementations of state-of-the-art research models spanning a wide range of modalities. Scenic has been successfully used for numerous projects and published papers and continues serving as the library of choice for rapid prototyping and publication of new research ideas.

****

CNN Filter DB: An Empirical Investigation of Trained Convolutional Filters

Paul Gavrikov, Janis Keuper; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19066-19076

Currently, many theoretical as well as practically relevant questions towards the transferability and robustness of Convolutional Neural Networks (CNNs) remain unsolved. While ongoing research efforts are engaging these problems from various angles, in most computer vision related cases these approaches can be generalized to investigations of the effects of distribution shifts in image data. In this context, we propose to study the shifts in the learned weights of trained CNN models. Here we focus on the properties of the distributions of dominantly used 3x3 convolution filter kernels. We collected and publicly provide a dataset with over 1.4 billion filters from hundreds of trained CNNs, using a wide range of datasets, architectures, and vision tasks. In a first use case of the proposed dataset, we can show highly relevant properties of many publicly available pre-trained models for practical applications: I) We analyze distribution shifts (or the lack thereof) between trained filters along different axes of meta-parameters, like visual category of the dataset, task, architecture, or layer depth. Based on these results, we conclude that model pre-training can succeed on arbitrary datasets if they meet size and variance conditions. II) We show that many pre-trained models contain degenerated filters which make them less robust and less suitable for fine-tuning on target applications.

****

ScePT: Scene-Consistent, Policy-Based Trajectory Predictions for Planning

Yuxiao Chen, Boris Ivanovic, Marco Pavone; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17103-17112

Trajectory prediction is a critical functionality of autonomous systems that share environments with uncontrolled agents, one prominent example being self-driving vehicles. Currently, most prediction methods do not enforce scene consistency, i.e., there are a substantial amount of self-collisions between predicted trajectories of different agents in the scene. Moreover, many approaches generate individual trajectory predictions per agent instead of joint trajectory predictions of the whole scene, which makes downstream planning difficult. In this work, we present ScePT, a policy planning-based trajectory prediction model that generates accurate, scene-consistent trajectory predictions suitable for autonomous system motion planning. It explicitly enforces scene consistency and learns an agent interaction policy that can be used for conditional prediction. Experiments on multiple real-world pedestrians and autonomous vehicle datasets show that ScePT matches current state-of-the-art prediction accuracy with significantly improved scene consistency. We also demonstrate ScePT's ability to work with a downstream contingency planner.

****

Calibrating Deep Neural Networks by Pairwise Constraints

Jiacheng Cheng, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13709-13718

It is well known that deep neural networks (DNNs) produce poorly calibrated estimates of class-posterior probabilities. We hypothesize that this is due to the limited calibration supervision provided by the cross-entropy loss, which places all emphasis on the probability of the true class and mostly ignores the remaini

ng. We consider how each example can supervise all classes and show that the cal ibration of a C-way classification problem is equivalent to the calibration of C (C-1)/2 pairwise binary classification problems that can be derived from it. Thi s suggests the hypothesis that DNN calibration can be improved by providing cali bration supervision to all such binary problems. An implementation of this calib ration by pairwise constraints (CPC) is then proposed, based on two types of bin ary calibration constraints. This is finally shown to be implementable with a ve ry minimal increase in the complexity of cross-entropy training. Empirical evalu ations of the proposed CPC method across multiple datasets and DNN architectures demonstrate state-of-the-art calibration performance.
****

Deep Saliency Prior for Reducing Visual Distraction
Kfir Aberman, Junfeng He, Yossi Gandelsman, Inbar Mosseri, David E. Jacobs, Kai Kohlhoff, Yael Pritch, Michael Rubinstein; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19851-19860
Using only a model that was trained to predict where people look at images, and no additional training data, we can produce a range of powerful editing effects for reducing distraction in images. Given an image and a mask specifying the reg ion to edit, we backpropagate through a state-of-the-art saliency model to param eterize a differentiable editing operator, such that the saliency within the mas ked region is reduced. We demonstrate several operators, including: a recoloring operator, which learns to apply a color transform that camouflages and blends d istractors into their surroundings; a warping operator, which warps less salient image regions to cover distractors, gradually collapsing objects into themselve s and effectively removing them (an effect akin to inpainting); a GAN operator, which uses a semantic prior to fully replace image regions with plausible, less salient alternatives. The resulting effects are consistent with cognitive resear ch on the human visual system (e.g., since color mismatch is salient, the recolo ring operator learns to harmonize objects' colors with their surrounding to redu ce their saliency), and, importantly, are all achieved solely through the guidan ce of the pretrained saliency model. We present results on a variety of natural images and conduct a perceptual study to evaluate and validate the changes in vi ewers' eye-gaze between the original images and our edited results.
****

Efficient Large-Scale Localization by Global Instance Recognition
Fei Xue, Ignas Budvytis, Daniel Olmeda Reino, Roberto Cipolla; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 17348-17357
Hierarchical frameworks consisting of both coarse and fine localization are ofte n used as the standard pipeline for large-scale visual localization. Despite the ir promising performance in simple environments, they still suffer from low effi ciency and accuracy in large-scale scenes, especially under challenging conditio ns. In this paper, we propose an efficient and accurate large-scale localization framework based on the recognition of buildings, which are not only discriminat ive for coarse localization but also robust for fine localization. Specifically, we assign each building instance a global ID and perform pixel-wise recognition of these global instances in the localization process. For coarse localization, we employ an efficient reference search strategy to find candidates progressive ly from the local map observing recognized instances instead of the whole databa se. For fine localization, predicted labels are further used for instance-wise f eature detection and matching, allowing our model to focus on fewer but more rob ust keypoints for establishing correspondences. The experiments in long-term lar ge-scale localization datasets including Aachen and RobotCar-Seasons demonstrate that our method outperforms previous approaches consistently in terms of both e fficiency and accuracy.
****

Sign Language Video Retrieval With Free-Form Textual Queries
Amanda Duarte, Samuel Albanie, Xavier Giró-i-Nieto, Gül Varol; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 14094-14104

Systems that can efficiently search collections of sign language videos have been highlighted as a useful application of sign language technology. However, the problem of searching videos beyond individual keywords has received limited attention in the literature. To address this gap, in this work we introduce the task of sign language retrieval with free-form textual queries: given a written query (e.g. a sentence) and a large collection of sign language videos, the objective is to find the signing video that best matches the written query. We propose to tackle this task by learning cross-modal embeddings on the recently introduced large-scale How2Sign dataset of American Sign Language (ASL). We identify that a key bottleneck in the performance of the system is the quality of the sign video embedding which suffers from a scarcity of labelled training data. We, therefore, propose SPOT-ALIGN, a framework for interleaving iterative rounds of sign spotting and feature alignment to expand the scope and scale of available training data. We validate the effectiveness of SPOT-ALIGN for learning a robust sign video embedding through improvements in both sign recognition and the proposed video retrieval task.

****

Real-Time Object Detection for Streaming Perception

Jinrong Yang, Songtao Liu, Zeming Li, Xiaoping Li, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5385-5395

Autonomous driving requires the model to perceive the environment and (re)act within a low latency for safety. While past works ignore the inevitable changes in the environment after processing, streaming perception is proposed to jointly evaluate the latency and accuracy into a single metric for video online perception. In this paper, instead of searching trade-offs between accuracy and speed like previous works, we point out that endowing real-time models with the ability to predict the future is the key to dealing with this problem. We build a simple and effective framework for streaming perception. It equips a novel DualFlow Perception module (DFP), which includes dynamic and static flows to capture the moving trend and basic detection feature for streaming prediction. Further, we introduce a Trend-Aware Loss (TAL) combined with a trend factor to generate adaptive weights for objects with different moving speeds. Our simple method achieves competitive performance on Argoverse-HD dataset and improves the AP by 4.9% compared to the strong baseline, validating its effectiveness. Our code will be made available at https://github.com/yancie-yjr/StreamYOLO.

****

Simulated Adversarial Testing of Face Recognition Models

Nataniel Ruiz, Adam Kortylewski, Weichao Qiu, Cihang Xie, Sarah Adel Bargal, Alan Yuille, Stan Sclaroff; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4145-4155

Most machine learning models are validated and tested on fixed datasets. This can give an incomplete picture of the capabilities and weaknesses of the model. Such weaknesses can be revealed at test time in the real world. The risks involved in such failures can be loss of profits, loss of time or even loss of life in certain critical applications. In order to alleviate this issue, simulators can be controlled in a fine-grained manner using interpretable parameters to explore the semantic image manifold. In this work, we propose a framework for learning how to test machine learning algorithms using simulators in an adversarial manner in order to find weaknesses in the model before deploying it in critical scenarios. We apply this method in a face recognition setup. We show that certain weaknesses of models trained on real data can be discovered using simulated samples. Using our proposed method, we can find adversarial synthetic faces that fool contemporary face recognition models. This demonstrates the fact that these models have weaknesses that are not measured by commonly used validation datasets. We hypothesize that this type of adversarial examples are not isolated, but usually lie in connected spaces in the latent space of the simulator. We present a method to find these adversarial regions as opposed to the typical adversarial points found in the adversarial example literature.

****

VisualHow: Multimodal Problem Solving
Jinhui Yang, Xianyu Chen, Ming Jiang, Shi Chen, Louis Wang, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15627-15637
Recent progress in the interdisciplinary studies of computer vision (CV) and natural language processing (NLP) has enabled the development of intelligent systems that can describe what they see and answer questions accordingly. However, despite showing usefulness in performing these vision-language tasks, existing methods still struggle in understanding real-life problems (i.e., how to do something) and suggesting step-by-step guidance to solve them. With an overarching goal of developing intelligent systems to assist humans in various daily activities, we propose VisualHow, a free-form and open-ended research that focuses on understanding a real-life problem and deriving its solution by incorporating key components across multiple modalities. We develop a new dataset with 20,028 real-life problems and 102,933 steps that constitute their solutions, where each step consists of both a visual illustration and a textual description that guide the problem solving. To establish better understanding of problems and solutions, we also provide annotations of multimodal attention that localizes important components across modalities and solution graphs that encapsulate different steps in structured representations. These data and annotations enable a family of new vision-language tasks that solve real-life problems. Through extensive experiments with representative models, we demonstrate their effectiveness on training and testing models for the new tasks, and there is significant scope for improvement by learning effective attention mechanisms. Our dataset and models are available at https://github.com/formidify/VisualHow.
****
Equivariance Allows Handling Multiple Nuisance Variables When Analyzing Pooled Neuroimaging Datasets
Vishnu Suresh Lokhande, Rudrasis Chakraborty, Sathya N. Ravi, Vikas Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10432-10441
Pooling multiple neuroimaging datasets across institutions often enables significant improvements in statistical power when evaluating associations (e.g., between risk factors and disease outcomes) that would otherwise be too weak to detect. When there is only a single source of variability (e.g., different scanners), domain adaptation and matching the distributions of representations may suffice in many scenarios. But in the presence of more than one nuisance variable which concurrently influence the measurements, pooling datasets poses unique challenges, e.g., variations in the data can come from both the acquisition method as well as the demographics of participants (gender, age). Invariant representation learning, by itself, is ill-suited to fully model the data generation process. In this paper, we show how bringing recent results on equivariant representation learning (for studying symmetries in neural networks) together with simple use of classical results on causal inference provides an effective practical solution to this problem. In particular, we demonstrate how our model allows dealing with more than one nuisance variable under some assumptions and can enable (relatively) painless analysis of pooled scientific datasets in scenarios that would otherwise entail removing a large portion of the samples.
****
Spatial Commonsense Graph for Object Localisation in Partial Scenes
Francesco Giuliari, Geri Skenderi, Marco Cristani, Yiming Wang, Alessio Del Bue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19518-19527
We solve object localisation in partial scenes, a new problem of estimating the unknown position of an object (e.g. where is the bag?) given a partial 3D scan of a scene. The proposed solution is based on a novel scene graph model, the Spatial Commonsense Graph (SCG), where objects are the nodes and edges define pairwise distances between them, enriched by concept nodes and relationships from a commonsense knowledge base. This allows SCG to better generalise its spatial inference to unknown 3D scenes. The SCG is used to estimate the unknown position of t

he target object in two steps: first, we feed the SCG into a novel Proximity Pre diction Network, a graph neural network that uses attention to perform distance prediction between the node representing the target object and the nodes represe nting the observed objects in the SCG; second, we propose a Localisation Module based on circular intersection to estimate the object position using all the pre dicted pairwise distances in order to be independent of any reference system. We create a new dataset of partially reconstructed scenes to benchmark our method and baselines for object localisation in partial scenes, where our proposed meth od achieves the best localisation performance. Code and Dataset are available he re: https://github.com/IIT-PAVIS/SpatialCommonsenseGraph
****

CAT-Det: Contrastively Augmented Transformer for Multi-Modal 3D Object Detection
Yanan Zhang, Jiaxin Chen, Di Huang; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2022, pp. 908-917
In autonomous driving, LiDAR point-clouds and RGB images are two major data moda lities with complementary cues for 3D object detection. However, it is quite dif ficult to sufficiently use them, due to large inter-modal discrepancies. To addr ess this issue, we propose a novel framework, namely Contrastively Augmented Tra nsformer for multi-modal 3D object Detection (CAT-Det). Specifically, CAT-Det ad opts a two-stream structure consisting of a Pointformer (PT) branch, an Imagefor mer (IT) branch along with a Cross-Modal Transformer (CMT) module. PT, IT and CM T jointly encode intra-modal and inter-modal long-range contexts for representin g an object, thus fully exploring multi-modal information for detection. Further more, we propose an effective One-way Multi-modal Data Augmentation (OMDA) appro ach via hierarchical contrastive learning at both the point and object levels, s ignificantly improving the accuracy only by augmenting point-clouds, which is fr ee from complex generation of paired samples of the two modalities. Extensive ex periments on the KITTI benchmark show that CAT-Det achieves a new state-of-the-a rt, highlighting its effectiveness.
****

OSSGAN: Open-Set Semi-Supervised Image Generation
Kai Katsumata, Duc Minh Vo, Hideki Nakayama; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11185-11193
We introduce a challenging training scheme of conditional GANs, called open-set semi-supervised image generation, where the training dataset consists of two par ts: (i) labeled data and (ii) unlabeled data with samples belonging to one of th e labeled data classes, namely, a closed-set, and samples not belonging to any o f the labeled data classes, namely, an open-set. Unlike the existing semi-superv ised image generation task, where unlabeled data only contain closed-set samples , our task is more general and lowers the data collection cost in practice by al lowing open-set samples to appear. Thanks to entropy regularization, the classif ier that is trained on labeled data is able to quantify sample-wise importance t o the training of cGAN as confidence, allowing us to use all samples in unlabele d data. We design OSSGAN, which provides decision clues to the discriminator on the basis of whether an unlabeled image belongs to one or none of the classes of interest, smoothly integrating labeled and unlabeled data during training. The results of experiments on Tiny ImageNet and ImageNet show notable improvements o ver supervised BigGAN and semisupervised methods. The code is available at https ://github.com/raven38/OSSGAN.
****

Lite Vision Transformer With Enhanced Self-Attention
Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, Alan Yu ille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2022, pp. 11998-12008
Despite the impressive representation capacity of vision transformer models, cur rent light-weight vision transformer models still suffer from inconsistent and i ncorrect dense predictions at local regions. We suspect that the power of their self-attention mechanism is limited in shallower and thinner networks. We propos e Lite Vision Transformer (LVT), a novel light-weight transformer network with t wo enhanced self-attention mechanisms to improve the model performances for mobi

le deployment. For the low-level features, we introduce Convolutional Self-Atten tion (CSA). Unlike previous approaches of merging convolution and self-attention , CSA introduces local self-attention into the convolution within a kernel of si ze 3 x 3 to enrich low-level features in the first stage of LVT. For the high-le vel features, we propose Recursive Atrous Self-Attention (RASA), which utilizes the multi-scale context when calculating the similarity map and a recursive mech anism to increase the representation capability with marginal extra parameter co st. The superiority of LVT is demonstrated on ImageNet recognition, ADE20K seman tic segmentation, and COCO panoptic segmentation. The code is made publicly avai lable.
****

Diversity Matters: Fully Exploiting Depth Clues for Reliable Monocular 3D Object Detection
Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, Lihui Jiang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 2791-2800
As an inherently ill-posed problem, depth estimation from single images is the m ost challenging part of monocular 3D object detection (M3OD). Many existing meth ods rely on preconceived assumptions to bridge the missing spatial information i n monocular images, and predict a sole depth value for every object of interest. However, these assumptions do not always hold in practical applications. To tac kle this problem, we propose a depth solving system that fully explores the visu al clues from the subtasks in M3OD and generates multiple estimations for the de pth of each target. Since the depth estimations rely on different assumptions in essence, they present diverse distributions. Even if some assumptions collapse, the estimations established on the remaining assumptions are still reliable. In addition, we develop a depth selection and combination strategy. This strategy is able to remove abnormal estimations caused by collapsed assumptions, and adap tively combine the remaining estimations into a single one. In this way, our dep th solving system becomes more precise and robust. Exploiting the clues from mul tiple subtasks of M3OD and without introducing any extra information, our method surpasses the current best method by more than 20% relatively on the Moderate l evel of test split in the KITTI 3D object detection benchmark, while still maint aining real-time efficiency.
****

NinjaDesc: Content-Concealing Visual Descriptors via Adversarial Learning
Tony Ng, Hyo Jin Kim, Vincent T. Lee, Daniel DeTone, Tsun-Yi Yang, Tianwei Shen, Eddy Ilg, Vassileios Balntas, Krystian Mikolajczyk, Chris Sweeney; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 12797-12807
In the light of recent analyses on privacy-concerning scene revelation from visu al descriptors, we develop descriptors that conceal the input image content. In particular, we propose an adversarial learning framework for training visual des criptors that prevent image reconstruction, while maintaining the matching accur acy. We let a feature encoding network and image reconstruction network compete with each other, such that the feature encoder tries to impede the image reconst ruction with its generated descriptors, while the reconstructor tries to recover the input image from the descriptors. The experimental results demonstrate that the visual descriptors obtained with our method significantly deteriorate the i mage reconstruction quality with minimal impact on correspondence matching and c amera localization performance.
****

Physically-Guided Disentangled Implicit Rendering for 3D Face Modeling
Zhenyu Zhang, Yanhao Ge, Ying Tai, Weijian Cao, Renwang Chen, Kunlin Liu, Hao Ta ng, Xiaoming Huang, Chengjie Wang, Zhifeng Xie, Dongjin Huang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 20353-20363
This paper presents a novel Physically-guided Disentangled Implicit Rendering (P hyDIR) framework for high-fidelity 3D face modeling. The motivation comes from t wo observations: widely-used graphics renderers yield excessive approximations a

gainst photo-realistic imaging, while neural rendering methods are highly entang
led to perceive 3D-aware operations. Hence, we learn to disentangle the implicit
 rendering via explicit physical guidance, meanwhile guarantee the properties of
 (1) 3D-aware comprehension and (2) high-reality imaging. For the former one, Ph
yDIR explicitly adopts 3D shading and rasterizing modules to control the rendere
r, which disentangles the lighting, facial shape and view point from neural reas
oning. Specifically, PhyDIR proposes a novel multi-image shading strategy to com
pensate the monocular limitation, so that the lighting variations are accessible
 to the neural renderer. For the latter one, PhyDIR learns the face-collection i
mplicit texture to avoid ill-posed intrinsic factorization, then leverages a ser
ies of consistency losses to constrain the robustness. With the disentangled met
hod, we make 3D face modeling benefit from both kinds of rendering strategies. E
xtensive experiments on benchmarks show that PhyDIR obtains superior performance
 than state-of-the-art explicit/implicit methods, on both geometry/texture model
ing.
****

M5Product: Self-Harmonized Contrastive Learning for E-Commercial Multi-Modal Pre
training

Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C. Kampffmeyer, Xiaoyon
g Wei, Minlong Lu, Yaowei Wang, Xiaodan Liang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21252-21262

Despite the potential of multi-modal pre-training to learn highly discriminative
 feature representations from complementary data modalities, current progress is
 being slowed by the lack of large-scale modality-diverse datasets. By leveragin
g the natural suitability of E-commerce, where different modalities capture comp
lementary semantic information, we contribute a large-scale multi-modal pre-trai
ning dataset M5Product. The dataset comprises 5 modalities (image, text, table,
video, and audio), covers over 6,000 categories and 5,000 attributes, and is 500
 times larger than the largest publicly available dataset with a similar number
of modalities. Furthermore, M5Product contains incomplete modality pairs and noi
se while also having a long-tailed distribution, resembling most real-world prob
lems. We further propose Self-harmonized ContrAstive LEarning (SCALE), a novel p
retraining framework that integrates the different modalities into a unified mod
el through an adaptive feature fusion mechanism, where the importance of each mo
dality is learned directly from the modality embeddings and impacts the inter-mo
dality contrastive learning and masked tasks within a multi-modal transformer mo
del. We evaluate the current multi-modal pre-training state-of-the-art approache
s and benchmark their ability to learn from unlabeled data when faced with the l
arge number of modalities in the M5Product dataset. We conduct extensive experim
ents on four downstream tasks and demonstrate the superiority of our SCALE model
, providing insights into the importance of dataset scale and diversity.
****

Bi-Level Alignment for Cross-Domain Crowd Counting

Shenjian Gong, Shanshan Zhang, Jian Yang, Dengxin Dai, Bernt Schiele; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2022, pp. 7542-7550

Recently, crowd density estimation has received increasing attention. The main c
hallenge for this task is to achieve high-quality manual annotations on a large
amount of training data. To avoid reliance on such annotations, previous works a
pply unsupervised domain adaptation (UDA) techniques by transferring knowledge l
earned from easily accessible synthetic data to real-world datasets. However, cu
rrent state-of-the-art methods either rely on external data for training an auxi
liary task or apply an expensive coarse-to-fine estimation. In this work, we aim
 to develop a new adversarial learning based method, which is simple and efficie
nt to apply. To reduce the domain gap between the synthetic and real data, we de
sign a bi-level alignment framework (BLA) consisting of (1) task-driven data ali
gnment and (2) fine-grained feature alignment. Contrast to previous domain augme
ntation methods, we introduce AutoML to search for an optimal transform on sourc
e, which well serves for the downstream task. On the other hand, we do fine-grai
ned alignment for foreground and background separately to alleviate the alignmen

t difficulty. We evaluate our approach on five real-world crowd counting benchma rks, where we outperform existing approaches by a large margin. Also, our approa ch is simple, easy to implement and efficient to apply. The code will be made pu blicly available. https://github.com/Yankeegsj/BLA
****

ST-MFNet: A Spatio-Temporal Multi-Flow Network for Frame Interpolation
Duolikun Danier, Fan Zhang, David Bull; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3521-3531
Video frame interpolation (VFI) is currently a very active research topic, with applications spanning computer vision, post production and video encoding. VFI c an be extremely challenging, particularly in sequences containing large motions, occlusions or dynamic textures, where existing approaches fail to offer percept ually robust interpolation performance. In this context, we present a novel deep learning based VFI method, ST-MFNet, based on a Spatio-Temporal Multi-Flow arch itecture. ST-MFNet employs a new multi-scale multi-flow predictor to estimate ma ny-to-one intermediate flows, which are combined with conventional one-to-one op tical flows to capture both large and complex motions. In order to enhance inter polation performance for various textures, a 3D CNN is also employed to model th e content dynamics over an extended temporal window. Moreover, ST-MFNet has been trained within an ST-GAN framework, which was originally developed for texture synthesis, with the aim of further improving perceptual interpolation quality. O ur approach has been comprehensively evaluated -- compared with fourteen state-o f-the-art VFI algorithms -- clearly demonstrating that ST-MFNet consistently out performs these benchmarks on varied and representative test datasets, with signi ficant gains up to 1.09dB in PSNR for cases including large motions and dynamic textures. Our source code is available at https://github.com/danielism97/ST-MFNe t.
****

Self-Supervised Super-Resolution for Multi-Exposure Push-Frame Satellites
Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, Gabriele Facciolo; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2022, pp. 1858-1868
Modern Earth observation satellites capture multi-exposure bursts of push-frame images that can be super-resolved via computational means. In this work, we prop ose a super-resolution method for such multi-exposure sequences, a problem that has received very little attention in the literature. The proposed method can ha ndle the signal-dependent noise in the inputs, process sequences of any length, and be robust to inaccuracies in the exposure times. Furthermore, it can be trai ned end-to-end with self-supervision, without requiring ground truth high resolu tion frames, which makes it especially suited to handle real data. Central to ou r method are three key contributions: i) a base-detail decomposition for handlin g errors in the exposure times, ii) a noise-level-aware feature encoding for imp roved fusion of frames with varying signal-to-noise ratio and iii) a permutation invariant fusion strategy by temporal pooling operators. We evaluate the propos ed method on synthetic and real data and show that it outperforms by a significa nt margin existing single-exposure approaches that we adapted to the multi-expos ure case.
****

Efficient Multi-View Stereo by Iterative Dynamic Cost Volume
Shaoqian Wang, Bo Li, Yuchao Dai; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2022, pp. 8655-8664
In this paper, we propose a novel iterative dynamic cost volume for multi-view s tereo. Compared with other works, our cost volume is much lighter, thus could be processed with 2D convolution based GRU. Notably, the every-step output of the GRU could be further used to generate new cost volume. In this way, an iterative GRU-based optimizer is constructed. Furthermore, we present a cascade and hiera rchical refinement architecture to utilize the multi-scale information and speed up the convergence. Specifically, a lightweight 3D CNN is utilized to generate the coarsest initial depth map which is essential to launch the GRU and guarante e a fast convergence. Then the depth map is refined by multi-stage GRUs which wo

rk on the pyramid feature maps. Extensive experiments on DTU and Tanks & Temples benchmarks demonstrate that our method could achieve state-of-the-art results in terms of accuracy, speed and memory usage. Code will be released at https://github.com/bdwsq1996/Effi-MVS.

****

Learning To Generate Line Drawings That Convey Geometry and Semantics

Caroline Chan, Frédo Durand, Phillip Isola; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7915-7925

This paper presents an unpaired method for creating line drawings from photographs. Current methods often rely on high quality paired datasets to generate line drawings. However, these datasets often have limitations due to the subjects of the drawings belonging to a specific domain, or in the amount of data collected. Although recent work in unsupervised image-to-image translation has shown much progress, the latest methods still struggle to generate compelling line drawings. We observe that line drawings are encodings of scene information and seek to convey 3D shape and semantic meaning. We build these observations into a set of objectives and train an image translation to map photographs into line drawings. We introduce a geometry loss which predicts depth information from the image features of a line drawing, and a semantic loss which matches the CLIP features of a line drawing with its corresponding photograph. Our approach outperforms state-of-the-art unpaired image translation and line drawing generation methods on creating line drawings from arbitrary photographs.

****

On Guiding Visual Attention With Language Specification

Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18092-18102

While real world challenges typically define visual categories with language words or phrases, most visual classification methods define categories with numerical indicies. However, the language specification of the classes provides an especially useful prior for biased and noisy datasets, where it can help disambiguate what features are task-relevant. Recently, large-scale multimodal models have been shown to recognize a wide variety of high-level concepts from a language specification even without additional image training data, but they are often unable to distinguish classes for more fine-grained tasks. CNNs, in contrast, can extract subtle image features that are required for fine-grained discrimination, but will overfit to any bias or noise in datasets. Our insight is to use high-level language specification as advice for constraining the prediction evidence to task-relevant features, instead of distractors. To do this, we ground task-relevant words or phrases with attention maps from a pretrained large-scale model. We then use this grounding to supervise a classifier's spatial attention away from distracting context. We show that supervising spatial attention in this way improves performance on classification tasks with biased and noisy data, including 3-15% worst-group accuracy improvements and 41-45% relative improvements on fairness metrics.

****

ReSTR: Convolution-Free Referring Image Segmentation Using Transformers

Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18145-18154

Referring image segmentation is an advanced semantic segmentation task where target is not a predefined class but is described in natural language. Most of existing methods for this task rely heavily on convolutional neural networks, which however have trouble capturing long-range dependencies between entities in the language expression and are not flexible enough for modeling interactions between the two different modalities. To address these issues, we present the first convolution-free model for referring image segmentation using transformers, dubbed ReSTR. Since it extracts features of both modalities through transformer encoders, it can capture long-range dependencies between entities within each modality. Also, ReSTR fuses features of the two modalities by a self-attention encoder, w

hich enables flexible and adaptive interactions between the two modalities in th
e fusion process. The fused features are fed to a segmentation module, which wor
ks adaptively according to the image and language expression in hand. ReSTR is e
valuated and compared with previous work on all public benchmarks, where it outp
erforms all existing models.
****
****
TransEditor: Transformer-Based Dual-Space GAN for Highly Controllable Facial Edi
ting
Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy,
Bo Dai, Wayne Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2022, pp. 7683-7692
Recent advances like StyleGAN have promoted the growth of controllable facial ed
iting. To address its core challenge of attribute decoupling in a single latent
space, attempts have been made to adopt dual-space GAN for better disentanglemen
t of style and content representations. Nonetheless, these methods are still inc
ompetent to obtain plausible editing results with high controllability, especial
ly for complicated attributes. In this study, we highlight the importance of int
eraction in a dual-space GAN for more controllable editing. We propose TransEdit
or, a novel Transformer-based framework to enhance such interaction. Besides, we
 develop a new dual-space editing and inversion strategy to provide additional e
diting flexibility. Extensive experiments demonstrate the superiority of the pro
posed framework in image quality and editing capability, suggesting the effectiv
eness of TransEditor for highly controllable facial editing. Code and models are
 publicly available at https://github.com/BillyXYB/TransEditor.
****
FLAG: Flow-Based 3D Avatar Generation From Sparse Observations
Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, Thomas J
. Cashman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2022, pp. 13253-13262
To represent people in mixed reality applications for collaboration and communic
ation, we need to generate realistic and faithful avatar poses. However, the sig
nal streams that can be applied for this task from head-mounted devices (HMDs) a
re typically limited to head pose and hand pose estimates. While these signals a
re valuable, they are an incomplete representation of the human body, making it
challenging to generate a faithful full-body avatar. We address this challenge b
y developing a flow-based generative model of the 3D human body from sparse obse
rvations, wherein we learn not only a conditional distribution of 3D human pose,
 but also a probabilistic mapping from observations to the latent space from whi
ch we can generate a plausible pose along with uncertainty estimates for the joi
nts. We show that our approach is not only a strong predictive model, but can al
so act as an efficient pose prior in different optimization settings where a goo
d initial latent code plays a major role.
****
Stability-Driven Contact Reconstruction From Monocular Color Images
Zimeng Zhao, Binghui Zuo, Wei Xie, Yangang Wang; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1643-1653
Physical contact provides additional constraints for hand-object state reconstru
ction as well as a basis for further understanding of interaction affordances. E
stimating these severely occluded regions from monocular images presents a consi
derable challenge. Existing methods optimize the hand-object contact driven by d
istance threshold or prior from contact-labeled datasets. However, due to the nu
mber of subjects and objects involved in these indoor datasets being limited, th
e learned contact patterns could not generalize easily. Our key idea is to recon
struct the contact pattern directly from monocular images and utilize the physic
al stability criterion in the simulation to drive the optimization process descr
ibed above. This criterion is defined by the resultant forces and contact distri
bution computed by the physics engine. Compared to existing solutions, our frame
work can be adapted to more personalized hands and diverse object shapes. Furthe
rmore, we create an interaction dataset with extra physical attributes to verify

the sim-to-real consistency of our methods. Through comprehensive evaluations, hand-object contact can be reconstructed with both accuracy and stability by the proposed framework.

****

Use All the Labels: A Hierarchical Multi-Label Contrastive Learning Framework
Shu Zhang, Ran Xu, Caiming Xiong, Chetan Ramaiah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16660-16669

Current contrastive learning frameworks focus on leveraging a single supervisory signal to learn representations, which limits the efficacy on unseen data and downstream tasks. In this paper, we present a hierarchical multi-label representation learning framework that can leverage all available labels and preserve the hierarchical relationship between classes. We introduce novel hierarchy preserving losses, which jointly apply a hierarchical penalty to the contrastive loss, and enforce the hierarchy constraint. The loss function is data driven and automatically adapts to arbitrary multi-label structures. Experiments on several datasets show that our relationship-preserving embedding performs well on a variety of tasks and outperform the baseline supervised and self-supervised approaches. Code is available at https://github.com/salesforce/hierarchicalContrastiveLearning.

****

SGTR: End-to-End Scene Graph Generation With Transformer
Rongjie Li, Songyang Zhang, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19486-19496

Scene Graph Generation (SGG) remains a challenging visual understanding task due to its compositional property. Most previous works adopt a bottom-up two-stage or a point-based one-stage approach, which often suffers from high time complexity or sub-optimal designs. In this work, we propose a novel SGG method to address the aforementioned issues, formulating the task as a bipartite graph construction problem. To solve the problem, we develop a transformer-based end-to-end framework that first generates the entity and predicate proposal set, followed by inferring directed edges to form the relation triplets. In particular, we develop a new entity-aware predicate representation based on a structural predicate generator that leverages the compositional property of relationships. Moreover, we design a graph assembling module to infer the connectivity of the bipartite scene graph based on our entity-aware structure, enabling us to generate the scene graph in an end-to-end manner. Extensive experimental results show that our design is able to achieve the state-of-the-art or comparable performance on two challenging benchmarks, surpassing most of the existing approaches and enjoying higher efficiency in inference. We hope our model can serve as a strong baseline for the Transformer-based scene graph generation. Code is available in https://github.com/Scarecrow0/SGTR

****

Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation
Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J. Guibas, Andrea Tagliasacchi, Frank Dellaert, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12871-12881

We present PanopticNeRF, an object-aware neural scene representation that decomposes a scene into a set of objects (things) and background (stuff). Each object is represented by a separate MLP that takes a position, direction, and time and outputs density and radiance. The background is represented by a similar MLP that also outputs semantics. Importantly, the object MLPs are specific to each instance and initialized with meta-learning, and thus can be smaller and faster than previous object-aware approaches, while still leveraging category-specific priors. We propose a system to infer the PanopticNeRF representation from a set of color images. We use off-the-shelf algorithms to predict camera poses, object bounding boxes, object categories, and 2D image semantic segmentations. Then we jointly optimize the MLP weights and bounding box parameters using analysis-by-synthesis with self-supervision from the color images and pseudo-supervision from pr

edicted semantic segmentations. PanopticNeRF can be effectively used for multipl
e 2D and 3D tasks like 3D scene editing, 3D panoptic reconstruction, novel view
and semantic synthesis, 2D panoptic segmentation, and multiview depth prediction
. We demonstrate these applications on several difficult, dynamic scenes with mo
ving objects.
****

Texture-Based Error Analysis for Image Super-Resolution
Salma Abdel Magid, Zudi Lin, Donglai Wei, Yulun Zhang, Jinjin Gu, Hanspeter Pfis
ter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2022, pp. 2118-2127
Evaluation practices for image super-resolution (SR) use a single-value metric,
the PSNR or SSIM, to determine model performance. This provides little insight i
nto the source of errors and model behavior. Therefore, it is beneficial to move
 beyond the conventional approach and reconceptualize evaluation with interpreta
bility as our main priority. We focus on a thorough error analysis from a variet
y of perspectives. Our key contribution is to leverage a texture classifier, whi
ch enables us to assign patches with semantic labels, to identify the source of
SR errors both globally and locally. We then use this to determine (a) the seman
tic alignment of SR datasets, (b) how SR models perform on each label, (c) to wh
at extent high-resolution (HR) and SR patches semantically correspond, and more.
 Through these different angles, we are able to highlight potential pitfalls and
 blindspots. Our overall investigation highlights numerous unexpected insights.
We hope this work serves as an initial step for debugging blackbox SR networks.
****

PILC: Practical Image Lossless Compression With an End-to-End GPU Oriented Neura
l Framework
Ning Kang, Shanzhao Qiu, Shifeng Zhang, Zhenguo Li, Shu-Tao Xia; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
 pp. 3739-3748
Generative model based image lossless compression algorithms have seen a great s
uccess in improving compression ratio. However, the throughput for most of them
is less than 1 MB/s even with the most advanced AI accelerated chips, preventing
 them from most real-world applications, which often require 100 MB/s. In this p
aper, we propose PILC, an end-to-end image lossless compression framework that a
chieves 200 MB/s for both compression and decompression with a single NVIDIA Tes
la V100 GPU, 10x faster than the most efficient one before. To obtain this resul
t, we first develop an AI codec that combines auto-regressive model and VQ-VAE w
hich performs well in lightweight setting, then we design a low complexity entro
py coder that works well with our codec. Experiments show that our framework com
presses better than PNG by a margin of 30% in multiple datasets. We believe this
 is an important step to bring AI compression forward to commercial use.
****

Set-Supervised Action Learning in Procedural Task Videos via Pairwise Order Cons
istency
Zijia Lu, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2022, pp. 19903-19913
We address the problem of set-supervised action learning, whose goal is to learn
 an action segmentation model using weak supervision in the form of sets of acti
ons occurring in training videos. Our key observation is that videos within the
same task have similar ordering of actions, which can be leveraged for effective
 learning. Therefore, we propose an attention-based method with a new Pairwise O
rdering Consistency (POC) loss that encourages that for each common action pair
in two videos of the same task, the attentions of actions follow a similar order
ing. Unlike existing sequence alignment methods, which misalign actions in video
s with different orderings or cannot reliably separate more from less consistent
 orderings, our POC loss efficiently aligns videos with different action orders
and is differentiable, which enables end-to-end training. In addition, it avoids
 the time-consuming pseudo-label generation of prior works. Our method efficient
ly learns the actions and their temporal locations, therefore, extends the exist
ing attention-based action localization methods from learning one action per vid

eo to multiple actions using our POC loss along with video-level and frame-level losses. By experiments on three datasets, we demonstrate that our method significantly improves the state of the art. We also show that our method, with a small modification, can effectively address the transcript-supervised action learning task, where actions and their ordering are available during training.
****

Learning To Align Sequential Actions in the Wild
Weizhe Liu, Bugra Tekin, Huseyin Coskun, Vibhav Vineet, Pascal Fua, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2181-2191
State-of-the-art methods for self-supervised sequential action alignment rely on deep networks that find correspondences across videos in time. They either learn frame-to-frame mapping across sequences, which does not leverage temporal information, or assume monotonic alignment between each video pair, which ignores variations in the order of actions. As such, these methods are not able to deal with common real-world scenarios that involve background frames or videos that contain non-monotonic sequence of actions. In this paper, we propose an approach to align sequential actions in the wild that involve diverse temporal variations. To this end, we propose an approach to enforce temporal priors on the optimal transport matrix, which leverages temporal consistency, while allowing for variations in the order of actions. Our model accounts for both monotonic and non-monotonic sequences and handles background frames that should not be aligned. We demonstrate that our approach consistently outperforms the state-of-the-art in self-supervised sequential action representation learning on four different benchmark datasets.
****

Decoupled Knowledge Distillation
Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, Jiajun Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11953-11962
State-of-the-art distillation methods are mainly based on distilling deep features from intermediate layers, while the significance of logit distillation is greatly overlooked. To provide a novel viewpoint to study logit distillation, we reformulate the classical KD loss into two parts, i.e., target class knowledge distillation (TCKD) and non-target class knowledge distillation (NCKD). We empirically investigate and prove the effects of the two parts: TCKD transfers knowledge concerning the "difficulty" of training samples, while NCKD is the prominent reason why logit distillation works. More importantly, we reveal that the classical KD loss is a coupled formulation, which (1) suppresses the effectiveness of NCKD and (2) limits the flexibility to balance these two parts. To address these issues, we present Decoupled Knowledge Distillation(DKD), enabling TCKD and NCKD to play their roles more efficiently and flexibly. Compared with complex feature-based methods, our DKD achieves comparable or even better results and has better training efficiency on CIFAR-100, ImageNet, and MS-COCO datasets for image classification and object detection tasks. This paper proves the great potential of logit distillation, and we hope it will be helpful for future research. The code is available at https://github.com/megvii-research/mdistiller.
****

DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection
Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, Mingxing Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17182-17191
Lidars and cameras are critical sensors that provide complementary information for 3D detection in autonomous driving. While prevalent multi-modal methods simply decorate raw lidar point clouds with camera features and feed them directly to existing 3D detection models, our study shows that fusing camera features with deep lidar features instead of raw points, can lead to better performance. However, as those features are often augmented and aggregated, a key challenge in fusion is how to effectively align the transformed features from two modalities. In

this paper, we propose two novel techniques: InverseAug that inverses geometric -related augmentations, e.g., rotation, to enable accurate geometric alignment b etween lidar points and image pixels, and LearnableAlign that leverages cross-at tention to dynamically capture the correlations between image and lidar features during fusion. Based on InverseAug and LearnableAlign, we develop a family of g eneric multi-modal 3D detection models named DeepFusion, which is more accurate than previous methods. For example, DeepFusion improves PointPillars, CenterPoin t, and 3D-MAN baselines on Pedestrian detection for 6.7, 8.9, and 6.2 LEVEL_2 AP H, respectively. Notably, our models achieve state-of-the-art performance on Way mo Open Dataset, and show strong model robustness against input corruptions and out-of-distribution data. Code will be publicly available at https://github.com/ tensorflow/lingvo.

****

Neural Volumetric Object Selection

Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G. Schwing, Oliver Wang ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 6133-6142

We introduce an approach for selecting objects in neural volumetric 3D represent ations, such as multi-plane images (MPI) and neural radiance fields (NeRF). Our approach takes a set of foreground and background 2D user scribbles in one view and automatically estimates a 3D segmentation of the desired object, which can b e rendered into novel views. To achieve this result, we propose a novel voxel fe ature embedding that incorporates the neural volumetric 3D representation and mu lti-view image features from all input views. To evaluate our approach, we intro duce a new dataset of human-provided segmentation masks for depicted objects in real-world multi-view scene captures. We show that our approach out-performs str ong baselines, including 2D segmentation and 3D segmentation approaches adapted to our task.

****

GCR: Gradient Coreset Based Replay Buffer Selection for Continual Learning

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, Pradeep Shenoy; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 99-108

Continual learning (CL) aims to develop techniques by which a single model adapt s to an increasing number of tasks encountered sequentially, thereby potentially leveraging learnings across tasks in a resource-efficient manner. A major chall enge for CL systems is catastrophic forgetting, where earlier tasks are forgotte n while learning a new task. To address this, replay-based CL approaches maintai n and repeatedly retrain on a small buffer of data selected across encountered t asks. We propose Gradient Coreset Replay (GCR), a novel strategy for replay buff er selection and update using a carefully designed optimization criterion. Speci fically, we select and maintain a 'coreset' that closely approximates the gradie nt of all the data seen so far with respect to current model parameters, and dis cuss key strategies needed for its effective application to the continual learni ng setting. We show significant gains (2%-4% absolute) over the state-of-the-art in the well-studied offline continual learning setting. Our findings also effec tively transfer to online / streaming CL settings, showing up to 5% gains over e xisting approaches. Finally, we demonstrate the value of supervised contrastive loss for continual learning, which yields a cumulative gain of up to 5% accuracy when combined with our subset selection strategy.

****

PointCLIP: Point Cloud Understanding by CLIP

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, P eng Gao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8552-8562

Recently, zero-shot and few-shot learning via Contrastive Vision-Language Pre-tr aining (CLIP) have shown inspirational performance on 2D visual recognition, whi ch learns to match images with their corresponding texts in open-vocabulary sett ings. However, it remains under explored that whether CLIP, pre-trained by large -scale image-text pairs in 2D, can be generalized to 3D recognition. In this pap

er, we identify such a setting is feasible by proposing PointCLIP, which conducts alignment between CLIP-encoded point cloud and 3D category texts. Specifically, we encode a point cloud by projecting it into multi-view depth maps without rendering, and aggregate the view-wise zero-shot prediction to achieve knowledge transfer from 2D to 3D. On top of that, we design an inter-view adapter to better extract the global feature and adaptively fuse the few-shot knowledge learned from 3D into CLIP pre-trained in 2D. By just fine-tuning the lightweight adapter in the few-shot settings, the performance of PointCLIP could be largely improved. In addition, we observe the complementary property between PointCLIP and classical 3D-supervised networks. By simple ensembling, PointCLIP boosts baseline's performance and even surpasses state-of-the-art models. Therefore, PointCLIP is a promising alternative for effective 3D point cloud understanding via CLIP under low resource cost and data regime. We conduct thorough experiments on widely-adopted ModelNet10, ModelNet40 and the challenging ScanObjectNN to demonstrate the effectiveness of PointCLIP.

****

NeRFusion: Fusing Radiance Fields for Large-Scale Scene Reconstruction

Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, Zexiang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5449-5458

While NeRF has shown great success for neural reconstruction and rendering, its limited MLP capacity and long per-scene optimization times make it challenging to model large-scale indoor scenes. In contrast, classical 3D reconstruction methods can handle large-scale scenes but do not produce realistic renderings. We propose NeRFusion, a method that combines the advantages of NeRF and TSDF-based fusion techniques to achieve efficient large-scale reconstruction and photo-realistic rendering. We process the input image sequence to predict per-frame local radiance fields via direct network inference. These are then fused using a novel recurrent neural network that incrementally reconstructs a global, sparse scene representation in real-time. This global volume can be further fine-tuned to boost rendering quality. We demonstrate that NeRFusion achieves state-of-the-art quality on both large-scale indoor and small-scale object scenes, with substantially faster reconstruction than NeRF and other recent methods.

****

DeepFace-EMD: Re-Ranking Using Patch-Wise Earth Mover's Distance Improves Out-of-Distribution Face Identification

Hai Phan, Anh Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20259-20269

Face identification (FI) is ubiquitous and drives many high-stake decisions made by the law enforcement. State-of-the-art FI approaches compare two images by taking the cosine similarity between their image embeddings. Yet, such approach suffers from poor out-of-distribution (OOD) generalization to new types of images (e.g., when a query face is masked, cropped or rotated) not included in the training set or the gallery. Here, we propose a re-ranking approach that compares two faces using the Earth Mover's Distance on the deep, spatial features of image patches. Our extra comparison stage explicitly examines image similarity at a fine-grained level (e.g., eyes to eyes) and is more robust to OOD perturbations and occlusions than traditional FI. Interestingly, without finetuning feature extractors, our method consistently improves the accuracy on all tested OOD queries: masked, cropped, rotated, and adversarial while obtaining similar results on in-distribution images.

****

A Sampling-Based Approach for Efficient Clustering in Large Datasets

Georgios Exarchakis, Omar Oubari, Gregor Lenz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12403-12412

We propose a simple and efficient clustering method for high-dimensional data with a large number of clusters. Our algorithm achieves high-performance by evaluating distances of datapoints with a subset of the cluster centres. Our contribution is substantially more efficient than k-means as it does not require an all to all comparison of data points and clusters. We show that the optimal solutions

of our approximation are the same as in the exact solution. However, our approa ch is considerably more efficient at extracting these clusters compared to the s tate-of-the-art. We compare our approximation with the exact k-means and alterna tive approximation approaches on a series of standardised clustering tasks. For the evaluation, we consider the algorithmic complexity, including the number of operations until convergence, and the stability of the results. An efficient imp lementation of the algorithm is provided in online.
****

General Facial Representation Learning in a Visual-Linguistic Manner
Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, L u Yuan, Dong Chen, Ming Zeng, Fang Wen; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18697-18709
How to learn a universal facial representation that boosts all face analysis tas ks This paper takes one step toward this goal. In this paper, we study the trans fer performance of pre-trained models on face analysis tasks and introduce a fra mework, called FaRL, for general facial representation learning. On one hand, th e framework involves a contrastive loss to learn high-level semantic meaning fro m image-text pairs. On the other hand, we propose exploring low-level informatio n simultaneously to further enhance the face representation by adding a masked i mage modeling. We perform pre-training on LAION-FACE, a dataset containing a lar ge amount of face image-text pairs, and evaluate the representation capability o n multiple downstream tasks. We show that FaRL achieves better transfer performa nce compared with previous pre-trained models. We also verify its superiority in the low-data regime. More importantly, our model surpasses the state-of-the-art methods on face analysis tasks including face parsing and face alignment.
****

Deep Color Consistent Network for Low-Light Image Enhancement
Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, Meng Wang; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1899-1908
Low-light image enhancement focus on refining the illumination and keep naturaln ess to obtain the normal-light image. Current low-light image enhancement method s can well improve the illumination. However, there is still color difference be tween the enhanced image and the ground-truth image. To alleviate this issue, we therefore propose deep color consistent network (DCC-Net) to preserve the color consistency for low-light enhancement. In this paper, we decouple a color image to two main components, a gray image and a color hist histogram. Further, we em ploy these two components to guide the enhancement, where the gray image is used to generate reasonable structures and textures and the corresponding color hist ogram is beneficial to keeping color consistency. To reduce the gap between imag es and color histograms, we also develop a pyramid color embedding (PCE) module, which can better embed the color information to the enhancement process accordi ng to the affinity between images and color histograms. Extensive experiments on the LOL, DICM, LIME, MEF, NPE and VV demonstrate that DCC-Net can well preserve color consistency, and performs favorably against state-of-the-art methods.
****

AdaSTE: An Adaptive Straight-Through Estimator To Train Binary Neural Networks
Huu Le, Rasmus Kjær Høier, Che-Tsung Lin, Christopher Zach; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 460-469
We propose a new algorithm for training deep neural networks (DNNs) with binary weights. In particular, we first cast the problem of training binary neural netw orks (BiNNs) as a bilevel optimization instance and subsequently construct flexi ble relaxations of this bilevel program. The resulting training method shares it s algorithmic simplicity with several existing approaches to train BiNNs, in par ticular with the straight-through gradient estimator successfully employed in Bi naryConnect and subsequent methods. In fact, our proposed method can be interpre ted as an adaptive variant of the original straight-through estimator that condi tionally (but not always) acts like a linear mapping in the backward pass of err or propagation. Experimental results demonstrate that our new algorithm offers f

avorable performance compared to existing approaches.
****

Reusing the Task-Specific Classifier as a Discriminator: Discriminator-Free Adversarial Domain Adaptation
Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, Enhong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7181-7190
Adversarial learning has achieved remarkable performances for unsupervised domain adaptation (UDA). Existing adversarial UDA methods typically adopt an additional discriminator to play the min-max game with a feature extractor. However, most of these methods failed to effectively leverage the predicted discriminative information, and thus cause mode collapse for generator. In this work, we address this problem from a different perspective and design a simple yet effective adversarial paradigm in the form of a discriminator-free adversarial learning network (DALN), wherein the category classifier is reused as a discriminator, which achieves explicit domain alignment and category distinguishment through a unified objective, enabling the DALN to leverage the predicted discriminative information for sufficient feature alignment. Basically, we introduce a Nuclear-norm Wasserstein discrepancy (NWD) that has definite guidance meaning for performing discrimination. Such NWD can be coupled with the classifier to serve as a discriminator satisfying the K-Lipschitz constraint without the requirements of additional weight clipping or gradient penalty strategy. Without bells and whistles, DALN compares favorably against the existing state-of-the-art (SOTA) methods on a variety of public datasets. Moreover, as a plug-and-play technique, NWD can be directly used as a generic regularizer to benefit existing UDA algorithms. Code is available at https://github.com/xiaoachen98/DALN.
****

Pooling Revisited: Your Receptive Field Is Suboptimal
Dong-Hwan Jang, Sanghyeok Chu, Joonhyuk Kim, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 549-558
The size and shape of the receptive field determine how the network aggregates local features, and affect the overall performance of a model considerably. Many components in a neural network, such as depth, kernel sizes, and strides for convolution and pooling, influence the receptive field. However, they still rely on hyperparameters, and the receptive fields of existing models result in suboptimal shapes and sizes. Hence, we propose a simple yet effective Dynamically Optimized Pooling operation, referred to as DynOPool, which learns the optimized scale factors of feature maps end-to-end. Moreover, DynOPool determines the proper resolution of a feature map by learning the desirable size and shape of its receptive field, which allows an operator in a deeper layer to observe an input image in the optimal scale. Any kind of resizing modules in a deep neural network can be replaced by DynOPool with minimal cost. Also, DynOPool controls the complexity of the model by introducing an additional loss term that constrains computational cost. Our experiments show that the models equipped with the proposed learnable resizing module outperform the baseline algorithms on multiple datasets in image classification and semantic segmentation.
****

Dual Task Learning by Leveraging Both Dense Correspondence and Mis-Correspondence for Robust Change Detection With Imperfect Matches
Jin-Man Park, Ue-Hwan Kim, Seon-Hoon Lee, Jong-Hwan Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13749-13759
Accurate change detection enables a wide range of tasks in visual surveillance, anomaly detection and mobile robotics. However, contemporary change detection approaches assume an ideal matching between the current and stored scenes, whereas only coarse matching is possible in real-world scenarios. Thus, contemporary approaches fail to show the reported performance in real-world settings. To overcome this limitation, we propose SimSaC. SimSaC concurrently conducts scene flow estimation and change detection and is able to detect changes with imperfect matc

hes. To train SimSaC without additional manual labeling, we propose a training s
cheme with random geometric transformations and the cut-paste method. Moreover,
we design an evaluation protocol which reflects performance in real-world settin
gs. In designing the protocol, we collect a test benchmark dataset, which we cla
im as another contribution. Our comprehensive experiments verify that SimSaC dis
plays robust performance even given imperfect matches and the performance margin
 compared to contemporary approaches is huge.
****

Show Me What and Tell Me How: Video Synthesis via Multimodal Conditioning
Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin
 Minaee, Dimitris Metaxas, Sergey Tulyakov; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3615-3625
Most methods for conditional video synthesis use a single modality as the condit
ion. This comes with major limitations. For example, it is problematic for a mod
el conditioned on an image to generate a specific motion trajectory desired by t
he user since there is no means to provide motion information. Conversely, langu
age information can describe the desired motion, while not precisely defining th
e content of the video. This work presents a multimodal video generation framewo
rk that benefits from text and images provided jointly or separately. We leverag
e the recent progress in quantized representations for videos and apply a bidire
ctional transformer with multiple modalities as inputs to predict a discrete vid
eo representation. To improve video quality and consistency, we propose a new vi
deo token trained with self-learning and an improved mask-prediction algorithm f
or sampling video tokens. We introduce text augmentation to improve the robustne
ss of the textual representation and diversity of generated videos. Our framewor
k can incorporate various visual modalities, such as segmentation masks, drawing
s, and partially occluded images. It can generate much longer sequences than the
 one used for training. In addition, our model can extract visual information as
 suggested by the text prompt, e.g., "an object in image one is moving northeast
", and generate corresponding videos. We run evaluations on three public dataset
s and a newly collected dataset labeled with facial attributes, achieving state-
of-the-art generation results on all four. [Code](https://github.com/snap-resear
ch/MMVID) and [webpage](https://snap-research.github.io/MMVID/).
****

Patch Slimming for Efficient Vision Transformers
Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, Dacheng Tao; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2022, pp. 12165-12174
This paper studies the efficiency problem for visual transformers by excavating
redundant calculation in given networks. The recent transformer architecture has
 demonstrated its effectiveness for achieving excellent performance on a series
of computer vision tasks. However, similar to that of convolutional neural netwo
rks, the huge computational cost of vision transformers is still a severe issue.
 Considering that the attention mechanism aggregates different patches layer-by-
layer, we present a novel patch slimming approach that discards useless patches
in a top-down paradigm. We first identify the effective patches in the last laye
r and then use them to guide the patch selection process of previous layers. For
 each layer, the impact of a patch on the final output feature is approximated a
nd patches with less impacts will be removed. Experimental results on benchmark
datasets demonstrate that the proposed method can significantly reduce the compu
tational costs of vision transformers without affecting their performances. For
example, over 45% FLOPs of the ViT-Ti model can be reduced with only 0.2% top-1
accuracy drop on the ImageNet dataset.
****

Bijective Mapping Network for Shadow Removal
Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, Zheng-Jun Zha; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 5627-5636
Shadow removal, which aims to restore the background in the shadow regions, is c
hallenging due to the highly ill-posed nature. Most existing deep learning-based

methods individually remove the shadow by only considering the content of the matched paired images, barely taking into account the auxiliary supervision of shadow generation on shadow removal procedure. In this work, we argue that shadow removal and generation are interrelated and could provide useful informative supervision for each other. Specifically, we propose a new Bijective Mapping Network (BMNet), which couples the learning procedures of shadow removal and shadow generation in a unified parameter-shared framework. With consistent two-way constraints and synchronous optimization of the two procedures, BMNet could effectively recover the underlying background contents during the forward shadow removal procedure. In addition, through statistic analysis on real-world datasets, we observe and verify that shadow appearances under different color spectrums are inconsistent. This motivates us to design a Shadow-Invariant Color Guidance Module (SICGM), which can explicitly utilize the learned shadow-invariant color information to guide network color restoration, thereby further reducing color-bias effects. Experiments on the representative ISTD, ISTD+ and SRD benchmarks show that our proposed network outperforms the state-of-the-art method [??] in de-shadowing performance, while only using its 0.25% network parameters and 6.25% floating point operations (FLOPs).

****

End-to-End Semi-Supervised Learning for Video Action Detection

Akash Kumar, Yogesh Singh Rawat; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14700-14710

In this work, we focus on semi-supervised learning for video action detection which utilizes both labeled as well as unlabeled data. We propose a simple end-to-end consistency based approach which effectively utilizes the unlabeled data. Video action detection requires both, action class prediction as well as a spatio-temporal localization of actions. Therefore, we investigate two types of constraints, classification consistency, and spatio-temporal consistency. The presence of predominant background and static regions in a video makes it challenging to utilize spatio-temporal consistency for action detection. To address this, we propose two novel regularization constraints for spatio-temporal consistency; 1) temporal coherency, and 2) gradient smoothness. Both these aspects exploit the temporal continuity of action in videos and are found to be effective for utilizing unlabeled videos for action detection. We demonstrate the effectiveness of the proposed approach on two different action detection benchmark datasets, UCF101-24 and JHMDB-21. In addition, we also show the effectiveness of the proposed approach for video object segmentation on the Youtube-VOS which demonstrates its generalization capability The proposed approach achieves competitive performance by using merely 20% of annotations on UCF101-24 when compared with recent fully supervised methods. On UCF101-24, it improves the score by +8.9% and +11% at 0.5 f-mAP and v-mAP respectively, compared to supervised approach.

****

Causal Transportability for Visual Recognition

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7521-7531

Visual representations underlie object recognition tasks, but they often contain both robust and non-robust features. Our main observation is that image classifiers may perform poorly on out-of-distribution samples because spurious correlations between non-robust features and labels can be changed in a new environment. By analyzing procedures for out-of-distribution generalization with a causal graph, we show that standard classifiers fail because the association between images and labels is not transportable across settings. However, we then show that the causal effect, which severs all sources of confounding, remains invariant across domains. This motivates us to develop an algorithm to estimate the causal effect for image classification, which is transportable (i.e., invariant) across source and target environments. Without observing additional variables, we show that we can derive an estimand for the causal effect under empirical assumptions using representations in deep models as proxies. Theoretical analysis, empirical results, and visualizations show that our approach captures causal invariances

and improves overall generalization.
****
Local Attention Pyramid for Scene Image Generation
Sang-Heon Shim, Sangeek Hyun, DaeHyun Bae, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7774-7782
In this paper, we first investigate the class-wise visual quality imbalance problem of scene images generated by GANs. The tendency is empirically found that the class-wise visual qualities are highly correlated with the dominance of object classes in the training data in terms of their scales and appearance frequencies. Specifically, the synthesized qualities of small and less frequent object classes tend to be low. To address this, we propose a novel attention module, Local Attention Pyramid (LAP) module tailored for scene image synthesis, that encourages GANs to generate diverse object classes in a high quality by explicit spread of high attention scores to local regions, since objects in scene images are scattered over the entire images. Moreover, our LAP assigns attention scores in a multiple scale to reflect the scale diversity of various objects. The experimental evaluations on three different datasets show consistent improvements in Frechet Inception Distance (FID) and Frechet Segmentation Distance (FSD) over the state-of-the-art baselines. Furthermore, we apply our LAP module to various GANs methods to demonstrate a wide applicability of our LAP module.
****
Multi-Objective Diverse Human Motion Prediction With Knowledge Distillation
Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, Chiho Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8161-8171
Obtaining accurate and diverse human motion prediction is essential to many industrial applications, especially robotics and autonomous driving. Recent research has explored several techniques to enhance diversity and maintain the accuracy of human motion prediction at the same time. However, most of them need to define a combined loss, such as the weighted sum of accuracy loss and diversity loss, and then decide their weights as hyperparameters before training. In this work, we aim to design a prediction framework that can balance the accuracy sampling and diversity sampling during the testing phase. In order to achieve this target, we propose a multi-objective conditional variational inference prediction model. We also propose a short-term oracle to encourage the prediction framework to explore more diverse future motions. We evaluate the performance of our proposed approach on two standard human motion datasets. The experiment results show that our approach is effective and on a par with state-of-the-art performance in terms of accuracy and diversity.
****
GridShift: A Faster Mode-Seeking Algorithm for Image Segmentation and Object Tracking
Abhishek Kumar, Oladayo S. Ajani, Swagatam Das, Rammohan Mallipeddi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8131-8139
In machine learning, MeanShift is one of the popular clustering algorithms. It iteratively moves each data point to the weighted mean of its neighborhood data points. The computational cost required for finding neighborhood data points for each one is quadratic to the number of data points. Therefore, it is very slow for large-scale datasets. To address this issue, we propose a mode-seeking algorithm, GridShift, with faster computing and principally based on MeanShift that uses a grid-based approach. To speed up, GridShift employs a grid-based approach for neighbor search, which is linear to the number of data points. In addition, GridShift moves the active grid cells (grid cells associated with at least one data point) in place of data points towards the higher density, which provides more speed up. The runtime of GridShift is linear to the number of active grid cells and exponential to the number of features. Therefore, it is ideal for large-scale low-dimensional applications such as object tracking and image segmentation. Through extensive experiments, we showcase the superior performance of GridShif

t compared to other MeanShift-based algorithms and state-of-the-art algorithms i
n terms of accuracy and runtime on benchmark datasets, image segmentation. Final
ly, we provide a new object-tracking algorithm based on GridShift and show promi
sing results for object tracking compared to camshift and MeanShift++.
****

Confidence Propagation Cluster: Unleash Full Potential of Object Detectors
Yichun Shen, Wanli Jiang, Zhen Xu, Rundong Li, Junghyun Kwon, Siyi Li; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 1151-1161
It's been a long history that most object detection methods obtain objects by us
ing the non-maximum suppression (NMS) and its improved versions like Soft-NMS to
 remove redundant bounding boxes. We challenge those NMS-based methods from thre
e aspects: 1) The bounding box with highest confidence value may not be the true
 positive having the biggest overlap with the ground-truth box. 2) Not only supp
ression is required for redundant boxes, but also confidence enhancement is need
ed for those true positives. 3) Sorting candidate boxes by confidence values is
not necessary so that full parallelism is achievable. In this paper, inspired by
 belief propagation (BP), we propose the Confidence Propagation Cluster (CP-Clus
ter) to replace NMS-based methods, which is fully parallelizable as well as bett
er in accuracy. In CP-Cluster, we borrow the message passing mechanism from BP t
o penalize redundant boxes and enhance true positives simultaneously in an itera
tive way until convergence. We verified the effectiveness of CP-Cluster by apply
ing it to various mainstream detectors such as FasterRCNN, SSD, FCOS, YOLOv3, YO
LOv5, Centernet etc. Experiments on MS COCO show that our plug and play method,
without retraining detectors, is able to steadily improve average mAP of all tho
se state-of-theart models with a clear margin from 0.3 to 1.9 respectively when
compared with NMS-based methods.
****

Cluster-Guided Image Synthesis With Unconditional Models
Markos Georgopoulos, James Oldfield, Grigorios G. Chrysos, Yannis Panagakis; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2022, pp. 11543-11552
Generative Adversarial Networks (GANs) are the driving force behind the state-of
-the-art in image generation. Despite their ability to synthesize high-resolutio
n photo-realistic images, generating content with on-demand conditioning of diff
erent granularity remains a challenge. This challenge is usually tackled by anno
tating massive datasets with the attributes of interest, a laborious task that i
s not always a viable option. Therefore, it is vital to introduce control into t
he generation process of unsupervised generative models. In this work, we focus
on controllable image generation by leveraging GANs that are well-trained in an
unsupervised fashion. To this end, we discover that the representation space of
intermediate layers of the generator forms a number of clusters that separate th
e data according to semantically meaningful attributes (e.g., hair color and pos
e). By conditioning on the cluster assignments, the proposed method is able to c
ontrol the semantic class of the generated image. Our approach enables sampling
from each cluster by Implicit Maximum Likelihood Estimation (IMLE). We showcase
the efficacy of our approach on faces (CelebA-HQ and FFHQ), animals (Imagenet) a
nd objects (LSUN) using different pre-trained generative models. The results hig
hlight the ability of our approach to condition image generation on attributes l
ike gender, pose and hair style on faces, as well as a variety of features on di
fferent object classes.
****

ISNet: Shape Matters for Infrared Small Target Detection
Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, Jie Guo; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2022, pp. 877-886
Infrared small target detection (IRSTD) refers to extracting small and dim targe
ts from blurred backgrounds, which has a wide range of applications such as traf
fic management and marine rescue. Due to the low signal-to-noise ratio and low c
ontrast, infrared targets are easily submerged in the background of heavy noise

and clutter. How to detect the precise shape information of infrared targets rem
ains challenging. In this paper, we propose a novel infrared shape network (ISNe
t), where Taylor finite difference (TFD)-inspired edge block and two-orientation
 attention aggregation (TOAA) block are devised to address this problem. Specifi
cally, TFD-inspired edge block aggregates and enhances the comprehensive edge in
formation from different levels, in order to improve the contrast between target
 and background and also lay a foundation for extracting shape information with
mathematical interpretation. TOAA block calculates the low-level information wit
h attention mechanism in both row and column directions and fuses it with the hi
gh-level information to capture the shape characteristic of targets and suppress
 noises. In addition, we construct a new benchmark consisting of 1,000 realistic
 images in various target shapes, different target sizes, and rich clutter backg
rounds with accurate pixel-level annotations, called IRSTD-1k. Experiments on pu
blic datasets and IRSTD-1k demonstrate the superiority of our approach over repr
esentative state-of-the-art IRSTD methods. The dataset and code are available at
 github.com/RuiZhang97/ISNet.
****

Robust Region Feature Synthesizer for Zero-Shot Object Detection
Peiliang Huang, Junwei Han, De Cheng, Dingwen Zhang; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7622-76
31
Zero-shot object detection aims at incorporating class semantic vectors to reali
ze the detection of (both seen and) unseen classes given an unconstrained test i
mage. In this study, we reveal the core challenges in this research area: how to
 synthesize robust region features (for unseen objects) that are as intra-class
diverse and inter-class separable as the real samples, so that strong unseen obj
ect detectors can be trained upon them. To address these challenges, we build a
novel zero-shot object detection framework that contains an Intra-class Semantic
 Diverging component and an Inter-class Structure Preserving component. The form
er is used to realize the one-to-more mapping to obtain diverse visual features
from each class semantic vector, preventing miss-classifying the real unseen obj
ects as image backgrounds. While the latter is used to avoid the synthesized fea
tures too scattered to mix up the inter-class and foreground-background relation
ship. To demonstrate the effectiveness of the proposed approach, comprehensive e
xperiments on PASCAL VOC, COCO, and DIOR datasets are conducted. Notably, our ap
proach achieves the new state-of-the-art performance on PASCAL VOC and COCO and
it is the first study to carry out zero-shot object detection in remote sensing
imagery.
****

Virtual Correspondence: Humans as a Cue for Extreme-View Geometry
Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, Antonio Torralba; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2022, pp. 15924-15934
Recovering the spatial layout of the cameras and the geometry of the scene from
extreme-view images is a longstanding challenge in computer vision. Prevailing 3
D reconstruction algorithms often adopt the image matching paradigm and presume
that a portion of the scene is co-visible across images, yielding poor performan
ce when there is little overlap among inputs. In contrast, humans can associate
visible parts in one image to the corresponding invisible components in another
image via prior knowledge of the shapes. Inspired by this fact, we present a nov
el concept called virtual correspondences (VCs). VCs are a pair of pixels from t
wo images whose camera rays intersect in 3D. Similar to classic correspondences,
 VCs conform with epipolar geometry; unlike classic correspondences, VCs do not
need to be co-visible across views. Therefore VCs can be established and exploit
ed even if images do not overlap. We introduce a method to find virtual correspo
ndences based on humans in the scene. We showcase how VCs can be seamlessly inte
grated with classic bundle adjustment to recover camera poses across extreme vie
ws. Experiments show that our method significantly outperforms state-of-the-art
camera pose estimation methods in challenging scenarios and is comparable in the
 traditional densely captured setup. Our approach also unleashes the potential o

f multiple downstream tasks such as scene reconstruction from multi-view stereo and novel view synthesis in extreme-view scenarios.
****

Segment, Magnify and Reiterate: Detecting Camouflaged Objects the Hard Way
Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, Zhongxuan Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4713-4722
It is challenging to accurately detect camouflaged objects from their highly similar surroundings. Existing methods mainly leverage a single-stage detection fashion, while neglecting small objects with low-resolution fine edges requires more operations than the larger ones. To tackle camouflaged object detection (COD), we are inspired by humans attention coupled with the coarse-to-fine detection strategy, and thereby propose an iterative refinement framework, coined SegMaR, which integrates Segment, Magnify and Reiterate in a multi-stage detection fashion. Specifically, we design a new discriminative mask which makes the model attend on the fixation and edge regions. In addition, we leverage an attention-based sampler to magnify the object region progressively with no need of enlarging the image size. Extensive experiments show our SegMaR achieves remarkable and consistent improvements over other state-of-the-art methods. Especially, we surpass two competitive methods 7.4% and 20.0% respectively in average over standard evaluation metrics on small camouflaged objects. Additional studies provide more promising insights into SegMaR, including its effectiveness on the discriminative mask and its generalization to other network architectures.
****

SIMBAR: Single Image-Based Scene Relighting for Effective Data Augmentation for Automated Driving Vision Tasks
Xianling Zhang, Nathan Tseng, Ameerah Syed, Rohan Bhasin, Nikita Jaipuria; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3718-3728
Real-world autonomous driving datasets comprise of images aggregated from different drives on the road. The ability to relight captured scenes to unseen lighting conditions, in a controllable manner, presents an opportunity to augment datasets with a richer variety of lighting conditions, similar to what would be encountered in the real-world. This paper presents a novel image-based relighting pipeline, SIMBAR, that can work with a single image as input. To the best of our knowledge, there is no prior work on scene relighting leveraging explicit geometric representations from a single image. We present qualitative comparisons with prior multi-view scene relighting baselines. To further validate and effectively quantify the benefit of leveraging SIMBAR for data augmentation for automated driving vision tasks, object detection and tracking experiments are conducted with a state-of-the-art method, a Multiple Object Tracking Accuracy (MOTA) of 93.3% is achieved with CenterTrack on SIMBAR-augmented KITTI - an impressive 9.0% relative improvement over the baseline MOTA of 85.6% with CenterTrack on original KITTI, both models trained from scratch and tested on Virtual KITTI. For more details and SIMBAR relit datasets, please visit our project website (https://simbarv1.github.io/).
****

Shape From Thermal Radiation: Passive Ranging Using Multi-Spectral LWIR Measurements
Yasuto Nagase, Takahiro Kushida, Kenichiro Tanaka, Takuya Funatomi, Yasuhiro Mukaigawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12661-12671
In this paper, we propose a new cue of depth sensing using thermal radiation. Our method realizes passive, texture independent, far range, and dark scene applicability, which can broaden the depth sensing subjects. A key observation is that thermal radiation is attenuated by the air and is wavelength dependent. By modeling the wavelength-dependent attenuation by the air and building a multi-spectral LWIR measurement system, we can jointly estimate the depth, temperature, and emissivity of the target. We analytically show the capability of the thermal radiation cue and show the effectiveness of the method in real-world scenes using a

n imaging system with a few bandpass filters.
****
Multi-Label Classification With Partial Annotations Using Class-Aware Selective Loss
Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, Lihi Zelnik-Manor; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4764-4772
Large-scale multi-label classification datasets are commonly, and perhaps inevitably, partially annotated. That is, only a small subset of labels are annotated per sample. Different methods for handling the missing labels induce different properties on the model and impact its accuracy. In this work, we analyze the partial labeling problem, then propose a solution based on two key ideas. First, un-annotated labels should be treated selectively according to two probability quantities: the class distribution in the overall dataset and the specific label likelihood for a given data sample. We propose to estimate the class distribution using a dedicated temporary model, and we show its improved efficiency over a naive estimation computed using the dataset's partial annotations. Second, during the training of the target model, we emphasize the contribution of annotated labels over originally un-annotated labels by using a dedicated asymmetric loss. With our novel approach, we achieve state-of-the-art results on OpenImages dataset (e.g. reaching 87.3 mAP on V6). In addition, experiments conducted on LVIS and simulated-COCO demonstrate the effectiveness of our approach. Code is available at https://github.com/Alibaba-MIIL/PartialLabelingCSL.
****
HSC4D: Human-Centered 4D Scene Capture in Large-Scale Indoor-Outdoor Space Using Wearable IMUs and LiDAR
Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6792-6802
We propose Human-centered 4D Scene Capture (HSC4D) to accurately and efficiently create a dynamic digital world, containing large-scale indoor-outdoor scenes, diverse human motions, and rich interactions between humans and environments. Using only body-mounted IMUs and LiDAR, HSC4D is space-free without any external devices' constraints and map-free without pre-built maps. Considering that IMUs can capture human poses but always drift for long-period use, while LiDAR is stable for global localization but rough for local positions and orientations, HSC4D makes both sensors complement each other by a joint optimization and achieves promising results for long-term capture. Relationships between humans and environments are also explored to make their interaction more realistic. To facilitate many down-stream tasks, like AR, VR, robots, autonomous driving, etc., we propose a dataset containing three large scenes (1k-5k m^2) with accurate dynamic human motions and locations. Diverse scenarios (climbing gym, multi-story building, slope, etc.) and challenging human activities (exercising, walking up/down stairs, climbing, etc.) demonstrate the effectiveness and the generalization ability of HSC4D. The dataset and code is available at lidarhumanmotion.net/hsc4d.
****
CADTransformer: Panoptic Symbol Spotting Transformer for CAD Drawings
Zhiwen Fan, Tianlong Chen, Peihao Wang, Zhangyang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10986-10996
Understanding 2D computer-aided design (CAD) drawings plays a crucial role for creating 3D prototypes in architecture, engineering and construction (AEC) industries. The task of automated panoptic symbol spotting, i.e., to spot and parse both countable object instances (windows, doors, tables, etc.) and uncountable stuff (wall, railing, etc.) from CAD drawings, has recently drawn interests from the computer vision community. Unfortunately, the highly irregular ordering and orientations set major roadblocks for this task. Existing methods, based on convolutional neural networks (CNNs) and/or graph neural networks (GNNs), regress instance bounding boxes in the pixel domain and then convert the predictions into symbols. In this paper, we present a novel framework named CADTransformer, that ca

n painlessly modify existing vision transformer (ViT) backbones to tackle the ab
ove limitations for the panoptic symbol spotting task. CADTransformer tokenizes
directly from the set of graphical primitives in CAD drawings, and corresponding
ly optimizes line-grained semantic and instance symbol spotting altogether by a
pair of prediction heads. The backbone is further enhanced with a few plug-and-p
lay modifications, including a neighborhood aware self-attention, hierarchical f
eature aggregation, and graphic entity position encoding, to bake in the structu
re prior while optimizing the efficiency. Besides, a new data augmentation metho
d, termed Random Layer, is proposed by the layer-wise separation and recombinati
on of a CAD drawing. Overall, CADTransformer significantly boosts the previous s
tate-of-the-art from 0.595 to 0.685 in the panoptic quality (PQ) metric, on the
recently released FloorPlanCAD dataset. We further demonstrate that our model ca
n spot symbols with irregular shapes and arbitrary orientations. Our codes are a
vailable in https://github.com/VITA-Group/CADTransformer.
****

IntraQ: Learning Synthetic Images With Intra-Class Heterogeneity for Zero-Shot N
etwork Quantization
Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghon
g Tian, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2022, pp. 12339-12348
Learning to synthesize data has emerged as a promising direction in zero-shot qu
antization (ZSQ), which represents neural networks by low-bit integer without ac
cessing any of the real data. In this paper, we observe an interesting phenomeno
n of intra-class heterogeneity in real data and show that existing methods fail
to retain this property in their synthetic images, which causes a limited perfor
mance increase. To address this issue, we propose a novel zero-shot quantization
 method referred to as IntraQ. First, we propose a local object reinforcement th
at locates the target objects at different scales and positions of the synthetic
 images. Second, we introduce a marginal distance constraint to form class-relat
ed features distributed in a coarse area. Lastly, we devise a soft inception los
s which injects a soft prior label to prevent the synthetic images from being ov
erfitting to a fixed object. Our IntraQ is demonstrated to well retain the intra
-class heterogeneity in the synthetic images and also observed to perform state-
of-the-art. For example, compared to the advanced ZSQ, our IntraQ obtains 9.17%
increase of the top-1 accuracy on ImageNet when all layers of MobileNetV1 are qu
antized to 4-bit. Code is at https://github.com/zysxmu/IntraQ.
****

M3L: Language-Based Video Editing via Multi-Modal Multi-Level Transformers
Tsu-Jui Fu, Xin Eric Wang, Scott T. Grafton, Miguel P. Eckstein, William Yang Wa
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2022, pp. 10513-10522
Video editing tools are widely used nowadays for digital design. Although the de
mand for these tools is high, the prior knowledge required makes it difficult fo
r novices to get started. Systems that could follow natural language instruction
s to perform automatic editing would significantly improve accessibility. This p
aper introduces the language-based video editing (LBVE) task, which allows the m
odel to edit, guided by text instruction, a source video into a target video. LB
VE contains two features: 1) the scenario of the source video is preserved inste
ad of generating a completely different video; 2) the semantic is presented diff
erently in the target video, and all changes are controlled by the given instruc
tion. We propose a Multi-Modal Multi-Level Transformer (M3L-Transformer) to carr
y out LBVE. The M3L-Transformer dynamically learns the correspondence between vi
deo perception and language semantic at different levels, which benefits both th
e video understanding and video frame synthesis. We build three new datasets for
 evaluation, including two diagnostic and one from natural videos with human-lab
eled text. Extensive experimental results show that M3L-Transformer is effective
 for video editing and that LBVE can lead to a new field toward vision-and-langu
age research.
****

I M Avatar: Implicit Morphable Head Avatars From Videos

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13545-13555

Traditional 3D morphable face models (3DMMs) provide fine-grained control over expression but cannot easily capture geometric and appearance details. Neural volumetric representations approach photorealism but are hard to animate and do not generalize well to unseen expressions. To tackle this problem, we propose IMavatar (Implicit Morphable avatar), a novel method for learning implicit head avatars from monocular videos. Inspired by the fine-grained control mechanisms afforded by conventional 3DMMs, we represent the expression- and pose- related deformations via learned blendshapes and skinning fields. These attributes are pose-independent and can be used to morph the canonical geometry and texture fields given novel expression and pose parameters. We employ ray marching and iterative root-finding to locate the canonical surface intersection for each pixel. A key contribution is our novel analytical gradient formulation that enables end- to-end training of IMavatars from videos. We show quantitatively and qualitatively that our method improves geometry and covers a more complete expression space compared to state-of-the-art methods. Code and data can be found at https://ait.ethz.ch/projects/2022/IMavatar/.

****

BodyMap: Learning Full-Body Dense Correspondence Map

Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, Tony Tung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13286-13295

Dense correspondence between humans carries powerful semantic information that can be utilized to solve fundamental problems for full-body understanding such as in-the-wild surface matching, tracking and reconstruction. In this paper we present BodyMap, a new framework for obtaining high-definition full-body and continuous dense correspondence between in-the-wild images of clothed humans and the surface of a 3D template model. The correspondences cover fine details such as hands and hair, while capturing regions far from the body surface, such as loose clothing. Prior methods for estimating such dense surface correspondence i) cut a 3D body into parts which are unwrapped to a 2D UV space, producing discontinuities along part seams, or ii) use a single surface for representing the whole body, but none handled body details. Here, we introduce a novel network architecture with Vision Transformers that learn fine-level features on a continuous body surface. BodyMap outperforms prior work on various metrics and datasets, including DensePose-COCO by a large margin. Furthermore, we show various applications ranging from multi-layer dense cloth correspondence, neural rendering with novel-view synthesis and appearance swapping.

****

Weakly-Supervised Metric Learning With Cross-Module Communications for the Classification of Anterior Chamber Angle Images

Jingqi Huang, Yue Ning, Dong Nie, Linan Guan, Xiping Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 752-762

As the basis for developing glaucoma treatment strategies, Anterior Chamber Angle (ACA) evaluation is usually dependent on experts' judgements. However, experienced ophthalmologists needed for these judgements are not widely available. Thus, computer-aided ACA evaluations become a pressing and efficient solution for this issue. In this paper, we propose a novel end-to-end framework GCNet for automated Glaucoma Classification based on ACA images or other Glaucoma-related medical images. We first collect and label an ACA image dataset with some pixel-level annotations. Next, we introduce a segmentation module and an embedding module to enhance the performance of classifying ACA images. Within GCNet, we design a Cross-Module Aggregation Net (CMANet) which is a weakly-supervised metric learning network to capture contextual information exchanging across these modules. We conduct experiments on the ACA dataset and two public datasets REFUGE and SIGF. Our experimental results demonstrate that GCNet outperforms several state-of-the-art deep models in the tasks of glaucoma medical image classifications. The sou

rce code of GCNet can be found at https://github.com/Jingqi-H/GCNet.
****
A Hybrid Egocentric Activity Anticipation Framework via Memory-Augmented Recurrent and One-Shot Representation Forecasting
Tianshan Liu, Kin-Man Lam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13904-13913
Egocentric activity anticipation involves identifying the interacted objects and target action patterns in the near future. A standard activity anticipation paradigm is recurrently forecasting future representations to compensate the missing activity semantics of the unobserved sequence. However, the limitations of current recursive prediction models arise from two aspects: (i) The vanilla recurrent units are prone to accumulated errors in relatively long periods of anticipation. (ii) The anticipated representations may be insufficient to reflect the desired semantics of the target activity, due to lack of contextual clues. To address these issues, we propose "HRO", a hybrid framework that integrates both the memory-augmented recurrent and one-shot representation forecasting strategies. Specifically, to solve the limitation (i), we introduce a memory-augmented contrastive learning paradigm to regulate the process of the recurrent representation forecasting. Since the external memory bank maintains long-term prototypical activity semantics, it can guarantee that the anticipated representations are reconstructed from the discriminative activity prototypes. To further guide the learning of the memory bank, two auxiliary loss functions are designed, based on the diversity and sparsity mechanisms, respectively. Furthermore, to resolve the limitation (ii), a one-shot transferring paradigm is proposed to enrich the forecasted representations, by distilling the holistic activity semantics after the target anticipation moment, in the offline training. Extensive experimental results on two large-scale data sets validate the effectiveness of our proposed HRO method.
****
It's All in the Teacher: Zero-Shot Quantization Brought Closer to the Teacher
Kanghyun Choi, Hye Yoon Lee, Deokki Hong, Joonsang Yu, Noseong Park, Youngsok Kim, Jinho Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8311-8321
Model quantization is considered as a promising method to greatly reduce the resource requirements of deep neural networks. To deal with the performance drop induced by quantization errors, a popular method is to use training data to fine-tune quantized networks. In real-world environments, however, such a method is frequently infeasible because training data is unavailable due to security, privacy, or confidentiality concerns. Zero-shot quantization addresses such problems, usually by taking information from the weights of a full-precision teacher network to compensate the performance drop of the quantized networks. In this paper, we first analyze the loss surface of state-of-the-art zero-shot quantization techniques and provide several findings. In contrast to usual knowledge distillation problems, zero-shot quantization often suffers from 1) the difficulty of optimizing multiple loss terms together, and 2) the poor generalization capability due to the use of synthetic samples. Furthermore, we observe that many weights fail to cross the rounding threshold during training the quantized networks even when it is necessary to do so for better performance. Based on the observations, we propose AIT, a simple yet powerful technique for zero-shot quantization, which addresses the aforementioned two problems in the following way: AIT i) uses a KL distance loss only without a cross-entropy loss, and ii) manipulates gradients to guarantee that a certain portion of weights are properly updated after crossing the rounding thresholds. Experiments show that AIT outperforms the performance of many existing methods by a great margin, taking over the overall state-of-the-art position in the field.
****
Improving Segmentation of the Inferior Alveolar Nerve Through Deep Label Propagation
Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, Costantino Grana; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter

n Recognition (CVPR), 2022, pp. 21137-21146
Many recent works in dentistry and maxillofacial imagery focused on the Inferior Alveolar Nerve (IAN) canal detection. Unfortunately, the small extent of available 3D maxillofacial datasets has strongly limited the performance of deep learning-based techniques. On the other hand, a huge amount of sparsely annotated data is produced every day from the regular procedures in the maxillofacial practice. Despite the amount of sparsely labeled images being significant, the adoption of those data still raises an open problem. Indeed, the deep learning approach frames the presence of dense annotations as a crucial factor. Recent efforts in literature have hence focused on developing label propagation techniques to expand sparse annotations into dense labels.However, the proposed methods proved only marginally effective for the purpose of segmenting the alveolar nerve in CBCT scans.This paper exploits and publicly releases a new 3D densely annotated dataset, through which we are able to train a deep label propagation model which obtains better results than those available in literature. By combining a segmentation model trained on the 3D annotated data and label propagation, we significantly improve the state of the art in the Inferior Alveolar Nerve segmentation.
****

A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-Resolution
Jianqi Ma, Zhetong Liang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5911-5920
Scene text image super-resolution aims to increase the resolution and readability of the text in low-resolution images. Though significant improvement has been achieved by deep convolutional neural networks (CNNs), it remains difficult to reconstruct high-resolution images for spatially deformed texts, especially rotated and curve-shaped ones. This is because the current CNN-based methods adopt locality-based operations, which are not effective to deal with the variation caused by deformations. In this paper, we propose a CNN based Text ATTention network (TATT) to address this problem. The semantics of the text are firstly extracted by a text recognition module as text prior information. Then we design a novel transformer-based module, which leverages global attention mechanism, to exert the semantic guidance of text prior to the text reconstruction process. In addition, we propose a text structure consistency loss to refine the visual appearance by imposing structural consistency on the reconstructions of regular and deformed texts. Experiments on the benchmark TextZoom dataset show that the proposed TATT not only achieves state-of-the-art performance in terms of PSNR/SSIM metrics, but also significantly improves the recognition accuracy in the downstream text recognition task, particularly for text instances with multi-orientation and curved shapes. Code is available at https://github.com/mjq11302010044/TATT.
****

Multi-Modal Dynamic Graph Transformer for Visual Grounding
Sijia Chen, Baochun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15534-15543
Visual grounding (VG) aims to align the correct regions of an image with a natural language query about that image. We found that existing VG methods are trapped by the single-stage grounding process that performs a sole evaluate-and- rank for meticulously prepared regions. Their performance depends on the density and quality of the candidate regions and is capped by the inability to optimize the located regions continuously. To address these issues, we propose to remodel VG into a progressively optimized visual semantic alignment process. Our proposed multi-modal dynamic graph transformer (M-DGT) achieves this by building upon the dynamic graph structure with regions as nodes and their semantic relations as edges. Starting from a few randomly initialized regions, M-DGT is able to make sustainable adjustments (i.e., 2D spatial transformation and deletion) to the nodes and edges of the graph based on multi-modal information and the graph feature, thereby efficiently shrinking the graph to approach the ground truth regions. Experiments show that with an average of 48 boxes as initialization, the performance of M-DGT on the Flickr30K entity and RefCOCO dataset outperforms existing state-of-the-art methods by a substantial margin in terms of both the accuracy and

Intersect over Union (IOU) scores. Furthermore, introducing M-DGT to optimize th
e predicted regions of existing methods can further significantly improve their
performance. The source codes are available at https://github.com/iQua/M-DGT.
****

OSOP: A Multi-Stage One Shot Object Pose Estimation Framework
Ivan Shugurov, Fu Li, Benjamin Busam, Slobodan Ilic; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6835-68
44
We present a novel one-shot method for object detection and 6 DoF pose estimatio
n, that does not require training on target objects. At test time, it takes as i
nput a target image and a textured 3D query model. The core idea is to represent
 a 3D model with a number of 2D templates rendered from different viewpoints. Th
is enables CNN-based direct dense feature extraction and matching. The object is
 first localized in 2D, then its approximate viewpoint is estimated, followed by
 dense 2D-3D correspondence prediction. The final pose is computed with PnP. We
evaluate the method on LineMOD, Occlusion, Homebrewed, YCB-V and TLESS datasets
and report very competitive performance in comparison to the state-of-the-art me
thods trained on synthetic data, even though our method is not trained on the ob
ject models used for testing.
****

Generative Cooperative Learning for Unsupervised Video Anomaly Detection
M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segu, Fisher Yu, Seung-Ik
 Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2022, pp. 14744-14754
Video anomaly detection is well investigated in weakly supervised and one-class
classification (OCC) settings. However, unsupervised video anomaly detection is
quite sparse, likely because anomalies are less frequent in occurrence and usual
ly not well-defined, which when coupled with the absence of ground truth supervi
sion, could adversely affect the convergence of learning algorithms. This proble
m is challenging yet rewarding as it can completely eradicate the costs of obtai
ning laborious annotations and enable such systems to be deployed without human
intervention. To this end, we propose a novel unsupervised Generative Cooperativ
e Learning (GCL) approach for video anomaly detection that exploits the low freq
uency of anomalies towards building a cross-supervision between a generator and
a discriminator. In essence, both networks get trained in a cooperative fashion,
 thereby facilitating the overall convergence. We conduct extensive experiments
on two large-scale video anomaly detection datasets, UCF crime and ShanghaiTech.
 Consistent improvement over the existing state-of-the-art unsupervised and OCC
methods corroborate the effectiveness of our approach.
****

Rethinking Semantic Segmentation: A Prototype View
Tianfei Zhou, Wenguan Wang, Ender Konukoglu, Luc Van Gool; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2
582-2593
Prevalent semantic segmentation solutions, despite their different network desig
ns (FCN based or attention based) and mask decoding strategies (parametric softm
ax based or pixel-query based), can be placed in one category, by considering th
e softmax weights or query vectors as learnable class prototypes. In light of th
is prototype view, this study uncovers several limitations of such parametric se
gmentation regime, and proposes a nonparametric alternative based on non-learnab
le prototypes. Instead of prior methods learning a single weight/query vector fo
r each class in a fully parametric manner, our model represents each class as a
set of non-learnable prototypes, relying solely on the mean features of several
training pixels within that class. The dense prediction is thus achieved by nonp
arametric nearest prototype retrieving. This allows our model to directly shape
the pixel embedding space, by optimizing the arrangement between embedded pixels
 and anchored prototypes. It is able to handle arbitrary number of classes with
a constant amount of learnable parameters.We empirically show that, with FCN bas
ed and attention based segmentation models (i.e., HRNet, Swin, SegFormer) and ba
ckbones (i.e., ResNet, HRNet, Swin, MiT), our nonparametric framework yields com

pelling results over several datasets (i.e., ADE20K, Cityscapes, COCO-Stuff), and performs well in the large-vocabulary situation. We expect this work will provoke a rethink of the current de facto semantic segmentation model design.
****

Geometric Transformer for Fast and Robust Point Cloud Registration
Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11143-11152
We study the problem of extracting accurate correspondences for point cloud registration. Recent keypoint-free methods bypass the detection of repeatable keypoints which is difficult in low-overlap scenarios, showing great potential in registration. They seek correspondences over downsampled superpoints, which are then propagated to dense points. Superpoints are matched based on whether their neighboring patches overlap. Such sparse and loose matching requires contextual features capturing the geometric structure of the point clouds. We propose Geometric Transformer to learn geometric feature for robust superpoint matching. It encodes pair-wise distances and triplet-wise angles, making it robust in low-overlap cases and invariant to rigid transformation. The simplistic design attains surprisingly high matching accuracy such that no RANSAC is required in the estimation of alignment transformation, leading to 100 times acceleration. Our method improves the inlier ratio by 17 30 percentage points and the registration recall by over 7 points on the challenging 3DLoMatch benchmark. Our code and models are available at https://github.com/qinzheng93/GeoTransformer.
****

Cross-Model Pseudo-Labeling for Semi-Supervised Action Recognition
Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, Stephen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2959-2968
Semi-supervised action recognition is a challenging but important task due to the high cost of data annotation. A common approach to this problem is to assign unlabeled data with pseudo-labels, which are then used as additional supervision in training. Typically in recent work, the pseudo-labels are obtained by training a model on the labeled data, and then using confident predictions from the model to teach itself.In this work, we propose a more effective pseudo-labeling scheme, called Cross-Model Pseudo-Labeling (CMPL). Concretely, we introduce a lightweight auxiliary network in addition to the primary backbone, and ask them to predict pseudo-labels for each other. We observe that, due to their different structural biases, these two models tend to learn complementary representations from the same video clips. Each model can thus benefit from its counterpart by utilizing cross-model predictions as supervision. Experiments on different data partition protocols demonstrate the significant improvement of our framework over existing alternatives. For example, CMPL achieves 7.6% and 25.1% Top-1 accuracy on Kinetics-400 and UCF-101 using only the RGB modality and 1% labeled data, outperforming our baseline model, FixMatch, by 9.0% and 10.3%, respectively.
****

UMT: Unified Multi-Modal Transformers for Joint Video Moment Retrieval and Highlight Detection
Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, Xiaohu Qie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3042-3051
Finding relevant moments and highlights in videos according to natural language queries is a natural and highly valuable common need in the current video content explosion era. Nevertheless, jointly conducting moment retrieval and highlight detection is an emerging research topic, even though its component problems and some related tasks have already been studied for a while. In this paper, we present the first unified framework, named Unified Multi-modal Transformers (UMT), capable of realizing such joint optimization while can also be easily degenerated for solving individual problems. As far as we are aware, this is the first scheme to integrate multi-modal (visual-audio) learning for either joint optimization or the individual moment retrieval task, and tackles moment retrieval as a ke

ypoint detection problem using a novel query generator and query decoder. Extens ive comparisons with existing methods and ablation studies on QVHighlights, Char ades-STA, YouTube Highlights, and TVSum datasets demonstrate the effectiveness, superiority, and flexibility of the proposed method under various settings. Sour ce code and pre-trained models are available at https://github.com/TencentARC/UM T.

****

Dual-Shutter Optical Vibration Sensing

Mark Sheinin, Dorian Chan, Matthew O'Toole, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16324-16333

Visual vibrometry is a highly useful tool for remote capture of audio, as well as the physical properties of materials, human heart rate, and more. While visually-observable vibrations can be captured directly with a high-speed camera, minute imperceptible object vibrations can be optically amplified by imaging the displacement of a speckle pattern, created by shining a laser beam on the vibrating surface. In this paper, we propose a novel method for sensing vibrations at high speeds (up to 63kHz), for multiple scene sources at once, using sensors rated for only 130Hz operation. Our method relies on simultaneously capturing the scene with two cameras equipped with rolling and global shutter sensors, respectively. The rolling shutter camera captures distorted speckle images that encode the highspeed object vibrations. The global shutter camera captures undistorted reference images of the speckle pattern, helping to decode the source vibrations. We demonstrate our method by capturing vibration caused by audio sources (e.g. speakers, human voice, and musical instruments) and analyzing the vibration modes of a tuning fork.

****

Demystifying the Neural Tangent Kernel From a Practical Perspective: Can It Be Trusted for Neural Architecture Search Without Training?

Jisoo Mok, Byunggook Na, Ji-Hoon Kim, Dongyoon Han, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11861-11870

In Neural Architecture Search (NAS), reducing the cost of architecture evaluation remains one of the most crucial challenges. Among a plethora of efforts to bypass training of each candidate architecture to convergence for evaluation, the Neural Tangent Kernel (NTK) is emerging as a promising theoretical framework that can be utilized to estimate the performance of a neural architecture at initialization. In this work, we revisit several at-initialization metrics that can be derived from the NTK and reveal their key shortcomings. Then, through the empirical analysis of the time evolution of NTK, we deduce that modern neural architectures exhibit highly non-linear characteristics, making the NTK-based metrics incapable of reliably estimating the performance of an architecture without some amount of training. To take such non-linear characteristics into account, we introduce Label-Gradient Alignment (LGA), a novel NTK-based metric whose inherent formulation allows it to capture the large amount of non-linear advantage present in modern neural architectures. With minimal amount of training, LGA obtains a meaningful level of rank correlation with the post-training test accuracy of an architecture. Lastly, we demonstrate that LGA, complemented with few epochs of training, successfully guides existing search algorithms to achieve competitive search performances with significantly less search cost. The code is available at: https://github.com/nutellamok/DemystifyingNTK.

****

Learning To Find Good Models in RANSAC

Daniel Barath, Luca Cavalli, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15744-15753

We propose the Model Quality Network, MQ-Net in short, for predicting the quality, e.g. the pose error of essential matrices, of models generated inside RANSAC. It replaces the traditionally used scoring techniques, e.g., inlier counting of RANSAC, truncated loss of MSAC, and the marginalization-based loss of MAGSAC++. Moreover, Minimal samples Filtering Network (MF-Net) is proposed for the early

rejection of minimal samples that likely lead to degenerate models or to ones th
at are inconsistent with the scene geometry, e.g., due to the chirality constrai
nt. We show on 54450 image pairs from public real-world datasets that the propos
ed MQ-Net leads to results superior to the state-of-the-art in terms of accuracy
 by a large margin. The proposed MF-Net accelerates the fundamental matrix estim
ation by five times and significantly reduces the essential matrix estimation ti
me while slightly improving accuracy as well. Also, we show experimentally that
consensus maximization, i.e. inlier counting, is not an inherently good measure
of the model quality for relative pose estimation. The code and models will be m
ade publicly available.
****

Interactiveness Field in Human-Object Interactions
Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, Chi-Keung Tang; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2022, pp. 20113-20122
Human-Object Interaction (HOI) detection plays a core role in activity understan
ding. Though recent two/one-stage methods have achieved impressive results, as a
n essential step, discovering interactive human-object pairs remains challenging
. Both one/two-stage methods fail to effectively extract interactive pairs inste
ad of generating redundant negative pairs. In this work, we introduce a previous
ly overlooked interactiveness bimodal prior: given an object in an image, after
pairing it with the humans, the generated pairs are either mostly non-interactiv
e, or mostly interactive, with the former more frequent than the latter. Based o
n this interactiveness bimodal prior we propose the "interactiveness field". To
make the learned field compatible with real HOI image considerations, we propose
 new energy constraints based on the cardinality and difference in the inherent
"interactiveness field" underlying interactive versus non-interactive pairs. Con
sequently, our method can detect more precise pairs and thus significantly boost
 HOI detection performance, which is validated on widely-used benchmarks where w
e achieve decent improvements overstate-of-the-arts. Our code will be made publi
cly available.
****

BodyGAN: General-Purpose Controllable Neural Human Body Generation
Chaojie Yang, Hanhui Li, Shengjie Wu, Shengkai Zhang, Haonan Yan, Nianhong Jiao,
 Jie Tang, Runnan Zhou, Xiaodan Liang, Tianxiang Zheng; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7733
-7742
Recent advances in generative adversarial networks (GANs) have provided potentia
l solutions for photorealistic human image synthesis. However, the explicit and
individual control of synthesis over multiple factors, such as poses, body shape
s, and skin colors, remains difficult for existing methods. This is because curr
ent methods mainly rely on a single pose/appearance model, which is limited in d
isentangling various poses and appearance in human images. In addition, such a u
nimodal strategy is prone to causing severe artifacts in the generated images li
ke color distortions and unrealistic textures. To tackle these issues, this pape
r proposes a multi-factor conditioned method dubbed BodyGAN. Specifically, given
 a source image, our Body-GAN aims at capturing the characteristics of the human
 body from multiple aspects: (i) A pose encoding branch consisting of three hybr
id subnetworks is adopted, to generate the semantic segmentation based represent
ation, the 3D surface based representation, and the key point based representati
on of the human body, respectively. (ii) Based on the segmentation results, an a
ppearance encoding branch is used to obtain the appearance information of the hu
man body parts. (iii) The outputs of these two branches are represented by user-
editable condition maps, which are then processed by a generator to predict the
synthesized image. In this way, our BodyGAN can achieve the fine-grained disenta
nglement of pose, body shape, and appearance, and consequently enable the explic
it and effective control of synthesis with diverse conditions. Extensive experim
ents on multiple datasets and a comprehensive user-study show that our BodyGAN a
chieves the state-of-the-art performance.
****

Image Disentanglement Autoencoder for Steganography Without Embedding
Xiyao Liu, Ziping Ma, Junxing Ma, Jian Zhang, Gerald Schaefer, Hui Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2303-2312
Conventional steganography approaches embed a secret message into a carrier for concealed communication but are prone to attack by recent advanced steganalysis tools. In this paper, we propose Image DisEntanglement Autoencoder for Steganography (IDEAS) as a novel steganography without embedding (SWE) technique. Instead of directly embedding the secret message into a carrier image, our approach hides it by transforming it into a synthesised image, and is thus fundamentally immune to typical steganalysis attacks. By disentangling an image into two representations for structure and texture, we exploit the stability of structure representation to improve secret message extraction while increasing synthesis diversity via randomising texture representations to enhance steganography security. In addition, we design an adaptive mapping mechanism to further enhance the diversity of synthesised images when ensuring different required extraction levels. Experimental results convincingly demonstrate IDEAS to achieve superior performance in terms of enhanced security, reliable secret message extraction and flexible adaptation for different extraction levels, compared to state-of-the-art SWE methods.
****

Self-Supervised Dense Consistency Regularization for Image-to-Image Translation
Minsu Ko, Eunju Cha, Sungjoo Suh, Huijin Lee, Jae-Joon Han, Jinwoo Shin, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18301-18310
Unsupervised image-to-image translation has gained considerable attention due to the recent impressive progress based on generative adversarial networks (GANs). In this paper, we present a simple but effective regularization technique for improving GAN-based image-to-image translation. To generate images with realistic local semantics and structures, we suggest to use an auxiliary self-supervised loss, enforcing point-wise consistency of the overlapped region between a pair of patches cropped from a single real image during training discriminators of GAN. Our experiment shows that the dense consistency regularization improves performance substantially on various image-to-image translation scenarios. It also achieves extra performance gains by using jointly with recent instance-level regularization methods. Furthermore, we verify that the proposed model captures domain-specific characteristics more effectively with only small fraction of training data.
****

The Devil Is in the Details: Window-Based Attention for Image Compression
Renjie Zou, Chunfeng Song, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17492-17501
Learned image compression methods have exhibited superior rate-distortion performance than classical image compression standards. Most existing learned image compression models are based on Convolutional Neural Networks (CNNs). Despite great contributions, a main drawback of CNN based model is that its structure is not designed for capturing local redundancy, especially the non-repetitive textures, which severely affects the reconstruction quality. Therefore, how to make full use of both global structure and local texture becomes the core problem for learning-based image compression. Inspired by recent progresses of Vision Transformer (ViT) and Swin Transformer, we found that combining the local-aware attention mechanism with the global-related feature learning could meet the expectation in image compression. In this paper, we first extensively study the effects of multiple kinds of attention mechanisms for local features learning, then introduce a more straightforward yet effective window-based local attention block. The proposed window-based attention is very flexible which could work as a plug-and-play component to enhance CNN and Transformer models. Moreover, we propose a novel Symmetrical TransFormer (STF) framework with absolute transformer blocks in the down-sampling encoder and up-sampling decoder. Extensive experimental evaluations have shown that the proposed method is effective and outperforms the state-of

-the-art methods. The code is publicly available at https://github.com/Googolxx/STF.

****

Category-Aware Transformer Network for Better Human-Object Interaction Detection
Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, Xu Zou ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19538-19547

Human-Object Interactions (HOI) detection, which aims to localize a human and a relevant object while recognizing their interaction, is crucial for understanding a still image. Recently, tranformer-based models have significantly advanced the progress of HOI detection. However, the capability of these models has not been fully explored since the Object Query of the model is always simply initialized as just zeros, which would affect the performance. In this paper, we try to study the issue of promoting transformerbased HOI detectors by initializing the Object Query with category-aware semantic information. To this end, we innovatively propose the Category-Aware Transformer Network (CATN). Specifically, the Object Query would be initialized via category priors represented by an external object detection model to yield a better performance. Moreover, such category priors can be further used for enhancing the representation ability of features via the attention mechanism. We have firstly verified our idea via the Oracle experiment by initializing the Object Query with the groundtruth category information. And then extensive experiments have been conducted to show that a HOI detection model equipped with our idea outperforms the baseline by a large margin to achieve a new state-of-the-art result.

****

Deep Depth From Focus With Differential Focus Volume
Fengting Yang, Xiaolei Huang, Zihan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12642-12651

Depth-from-focus (DFF) is a technique that infers depth using the focus change of a camera. In this work, we propose a convolutional neural network (CNN) to find the best-focused pixels in a focal stack and infer depth from the focus estimation. The key innovation of the network is the novel deep differential focus volume (DFV). By computing the first-order derivative with the stacked features over different focal distances, DFV is able to capture both the focus and context information for focus analysis. Besides, we also introduce a probability regression mechanism for focus estimation to handle sparsely sampled focal stacks and provide uncertainty estimation to the final prediction. Comprehensive experiments demonstrate that the proposed model achieves state-of-the-art performance on multiple datasets with good generalizability and fast speed.

****

DiLiGenT102: A Photometric Stereo Benchmark Dataset With Controlled Shape and Material Variation
Jieji Ren, Feishi Wang, Jiahao Zhang, Qian Zheng, Mingjun Ren, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12581-12590

Evaluating photometric stereo using real-world dataset is important yet difficult. Existing datasets are insufficient due to their limited scale and random distributions in shape and material. This paper presents a new real-world photometric stereo dataset with "ground truth" normal maps, which is 10 times larger than the widely adopted one. More importantly, we propose to control the shape and material variations by fabricating objects from CAD models with carefully selected materials, covering typical aspects of reflectance properties that are distinctive for evaluating photometric stereo methods. By benchmarking recent photometric stereo methods using these 100 sets of images, with a special focus on recent learning based solutions, a 10 x 10 shape-material error distribution matrix is visualized to depict a "portrait" for each evaluated method. From such comprehensive analysis, open problems in this field are discussed. To inspire future research, this dataset is available at https://photometricstereo.github.io.

****

Robust Fine-Tuning of Zero-Shot Models

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, Ludwig Schmidt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7959-7971

Large pre-trained models such as CLIP or ALIGN offer consistent accuracy across a range of data distributions when performing zero-shot inference (i.e., without fine-tuning on a specific dataset). Although existing fine-tuning methods substantially improve accuracy on a given target distribution, they often reduce robustness to distribution shifts. We address this tension by introducing a simple and effective method for improving robustness while fine-tuning: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements under distribution shift, while preserving high accuracy on the target distribution. On ImageNet and five derived distribution shifts, WiSE-FT improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while increasing ImageNet accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness gains (2 to 23 pp) on a diverse set of six further distribution shifts, and accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.

****

Towards Data-Free Model Stealing in a Hard Label Setting

Sunandini Sanyal, Sravanti Addepalli, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15284-15293

Machine learning models deployed as a service (MLaaS) are susceptible to model stealing attacks, where an adversary attempts to steal the model within a restricted access framework. While existing attacks demonstrate near-perfect clone-model performance using softmax predictions of the classification network, most of the APIs allow access to only the top-1 labels. In this work, we show that it is indeed possible to steal Machine Learning models by accessing only top-1 predictions (Hard Label setting) as well, without access to model gradients (Black-Box setting) or even the training dataset (Data-Free setting) within a low query budget. We propose a novel GAN-based framework that trains the student and generator in tandem to steal the model effectively while overcoming the challenge of the hard label setting by utilizing gradients of the clone network as a proxy to the victim's gradients. We propose to overcome the large query costs associated with a typical Data-Free setting by utilizing publicly available (potentially unrelated) datasets as a weak image prior. We additionally show that even in the absence of such data, it is possible to achieve state-of-the-art results within a low query budget using synthetically crafted samples. We are the first to demonstrate the scalability of Model Stealing in a restricted access setting on a 100 class dataset as well.

****

PolyWorld: Polygonal Building Extraction With Graph Neural Networks in Satellite Images

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, Friedrich Fraundorfer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1848-1857

While most state-of-the-art instance segmentation methods produce binary segmentation masks, geographic and cartographic applications typically require precise vector polygons of extracted objects instead of rasterized output. This paper introduces PolyWorld, a neural network that directly extracts building vertices from an image and connects them correctly to create precise polygons. The model predicts the connection strength between each pair of vertices using a graph neural network and estimates the assignments by solving a differentiable optimal transport problem. Moreover, the vertex positions are optimized by minimizing a combined segmentation and polygonal angle difference loss. PolyWorld significantly outperforms the state of the art in building polygonization and achieves not only notable quantitative results, but also produces visually pleasing building poly

gons. Code and trained weights are publicly available at https://github.com/zorzi-s/PolyWorldPretrainedNetwork.
****

GAT-CADNet: Graph Attention Network for Panoptic Symbol Spotting in CAD Drawings
Zhaohua Zheng, Jianfang Li, Lingjie Zhu, Honghua Li, Frank Petzold, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11747-11756
Spotting graphical symbols from the computer-aided design (CAD) drawings is essential to many industrial applications. Different from raster images, CAD drawings are vector graphics consisting of geometric primitives such as segments, arcs, and circles. By treating each CAD drawing as a graph, we propose a novel graph attention network GAT-CADNet to solve the panoptic symbol spotting problem: vertex features derived from the GAT branch are mapped to semantic labels, while their attention scores are cascaded and mapped to instance prediction. Our key contributions are three-fold: 1) the instance symbol spotting task is formulated as a subgraph detection problem and solved by predicting the adjacency matrix; 2) a relative spatial encoding (RSE) module explicitly encodes the relative positional and geometric relation among vertices to enhance the vertex attention; 3) a cascaded edge encoding (CEE) module extracts vertex attentions from multiple stages of GAT and treats them as edge encoding to predict the adjacency matrix. The proposed GAT-CADNet is intuitive yet effective and manages to solve the panoptic symbol spotting problem in one consolidated network. Extensive experiments and ablation studies on the public benchmark show that our graph-based approach surpasses existing state-of-the-art methods by a large margin.
****

Multi-Granularity Alignment Domain Adaptation for Object Detection
Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, Yanjun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9581-9590
Domain adaptive object detection is challenging due to distinctive data distribution between source domain and target domain. In this paper, we propose a unified multi-granularity alignment based object detection framework towards domain-invariant feature learning. To this end, we encode the dependencies across different granularity perspectives including pixel-, instance-, and category-levels simultaneously to align two domains. Based on pixel-level feature maps from the backbone network, we first develop the omni-scale gated fusion module to aggregate discriminative representations of instances by scale-aware convolutions, leading to robust multi-scale object detection. Meanwhile, the multi-granularity discriminators are proposed to identify which domain different granularities of samples (i.e., pixels, instances, and categories) come from. Notably, we leverage not only the instance discriminability in different categories but also the category consistency between two domains. Extensive experiments are carried out on multiple domain adaptation scenarios, demonstrating the effectiveness of our framework over state-of-the-art algorithms on top of anchor-free FCOS and anchor-based Faster R-CNN detectors with different backbones.
****

LARGE: Latent-Based Regression Through GAN Semantics
Yotam Nitzan, Rinon Gal, Ofir Brenner, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19239-19249
We propose a novel method for solving regression tasks using few-shot or weak supervision. At the core of our method is the fundamental observation that GANs are incredibly successful at encoding semantic information within their latent space, even in a completely unsupervised setting. For modern generative frameworks, this semantic encoding manifests as smooth, linear directions which affect image attributes in a disentangled manner. These directions have been widely used in GAN-based image editing. In this work, we leverage them for few-shot regression. Specifically, we make the simple observation that distances traversed along such directions are good features for downstream tasks - reliably gauging the magnitude of a property in an image. In the absence of explicit supervision, we use

these distances to solve tasks such as sorting a collection of images, and ordinal regression. With a few labels -- as little as two -- we calibrate these distances to real-world values and convert a pre-trained GAN into a state-of-the-art few-shot regression model. This enables solving regression tasks on datasets and attributes which are difficult to produce quality supervision for. Extensive experimental evaluations demonstrate that our method can be applied across a wide range of domains, leverage multiple latent direction discovery frameworks, and achieve state-of-the-art results in few-shot and low-supervision settings, even when compared to methods designed to tackle a single task. Code is available on our project website.
****

Are Multimodal Transformers Robust to Missing Modality?
Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18177-18186
Multimodal data collected from the real world are often imperfect due to missing modalities. Therefore multimodal models that are robust against modal-incomplete data are highly preferred. Recently, Transformer models have shown great success in processing multimodal data. However, existing work has been limited to either architecture designs or pre-training strategies; whether Transformer models are naturally robust against missing-modal data has rarely been investigated. In this paper, we present the first-of-its-kind work to comprehensively investigate the behavior of Transformers in the presence of modal-incomplete data. Unsurprising, we find Transformer models are sensitive to missing modalities while different modal fusion strategies will significantly affect the robustness. What surprised us is that the optimal fusion strategy is dataset dependent even for the same Transformer model; there does not exist a universal strategy that works in general cases. Based on these findings, we propose a principle method to improve the robustness of Transformer models by automatically searching for an optimal fusion strategy regarding input data. Experimental validations on three benchmarks support the superior performance of the proposed method.
****

Degradation-Agnostic Correspondence From Resolution-Asymmetric Stereo
Xihao Chen, Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Yueyi Zhang, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12962-12971
In this paper, we study the problem of stereo matching from a pair of images with different resolutions, e.g., those acquired with a tele-wide camera system. Due to the difficulty of obtaining ground-truth disparity labels in diverse real-world systems, we start from an unsupervised learning perspective. However, resolution asymmetry caused by unknown degradations between two views hinders the effectiveness of the generally assumed photometric consistency. To overcome this challenge, we propose to impose the consistency between two views in a feature space instead of the image space, named feature-metric consistency. Interestingly, we find that, although a stereo matching network trained with the photometric loss is not optimal, its feature extractor can produce degradation-agnostic and matching-specific features. These features can then be utilized to formulate a feature-metric loss to avoid the photometric inconsistency. Moreover, we introduce a self-boosting strategy to optimize the feature extractor progressively, which further strengthens the feature-metric consistency. Experiments on both simulated datasets with various degradations and a self-collected real-world dataset validate the superior performance of the proposed method over existing solutions.
****

Fisher Information Guidance for Learned Time-of-Flight Imaging
Jiaqu Li, Tao Yue, Sijie Zhao, Xuemei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16334-16343
Indirect Time-of-Flight (ToF) imaging is widely applied in practice for its superiorities on cost and spatial resolution. However, lower signal-to-noise ratio (SNR) of measurement leads to larger error in ToF imaging, especially for imaging scenes with strong ambient light or long distance. In this paper, we propose a

Fisher-information guided framework to jointly optimize the coding functions (light modulation and sensor demodulation functions) and the reconstruction network of iToF imaging, with the supervision of the proposed discriminative fisher loss. By introducing the differentiable modeling of physical imaging process considering various real factors and constraints, e.g., light-falloff with distance, physical implementability of coding functions, etc., followed by a dual-branch depth reconstruction neural network, the proposed method could learn the optimal iToF imaging system in an end-to-end manner. The effectiveness of the proposed method is extensively verified with both simulations and prototype experiments.
****

VRDFormer: End-to-End Video Visual Relation Detection With Transformers
Sipeng Zheng, Shizhe Chen, Qin Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18836-18846
Visual relation understanding plays an essential role for holistic video understanding. Most previous works adopt a multi-stage framework for video visual relation detection (VidVRD), which cannot capture long-term spatiotemporal contexts in different stages and also suffers from inefficiency. In this paper, we propose a transformerbased framework called VRDFormer to unify these decoupling stages. Our model exploits a query-based approach to autoregressively generate relation instances. We specifically design static queries and recurrent queries to enable efficient object pair tracking with spatio-temporal contexts. The model is jointly trained with object pair detection and relation classification. Extensive experiments on two benchmark datasets, ImageNet-VidVRD and VidOR, demonstrate the effectiveness of the proposed VRDFormer, which achieves the state-of-the-art performance on both relation detection and relation tagging tasks.
****

Robust Federated Learning With Noisy and Heterogeneous Clients
Xiuwen Fang, Mang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10072-10081
Model heterogeneous federated learning is a challenging task since each client independently designs its own model. Due to the annotation difficulty and free-riding participant issue, the local client usually contains unavoidable and varying noises, which cannot be effectively addressed by existing algorithms. This paper starts the first attempt to study a new and challenging robust federated learning problem with noisy and heterogeneous clients. We present a novel solution RHFL (Robust Heterogeneous Federated Learning), which simultaneously handles the label noise and performs federated learning in a single framework. It is featured in three aspects: (1) For the communication between heterogeneous models, we directly align the models feedback by utilizing public data, which does not require additional shared global models for collaboration. (2) For internal label noise, we apply a robust noise-tolerant loss function to reduce the negative effects. (3) For challenging noisy feedback from other participants, we design a novel client confidence re-weighting scheme, which adaptively assigns corresponding weights to each client in the collaborative learning stage. Extensive experiments validate the effectiveness of our approach in reducing the negative effects of different noise rates/types under both model homogeneous and heterogeneous federated learning settings, consistently outperforming existing methods.
****

Enabling Equivariance for Arbitrary Lie Groups
Lachlan E. MacDonald, Sameera Ramasinghe, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8183-8192
Although provably robust to translational perturbations, convolutional neural networks (CNNs) are known to suffer from extreme performance degradation when presented at test time with more general geometric transformations of inputs. Recently, this limitation has motivated a shift in focus from CNNs to Capsule Networks (CapsNets). However, CapsNets suffer from admitting relatively few theoretical guarantees of invariance. We introduce a rigourous mathematical framework to permit invariance to any Lie group of warps, exclusively using convolutions (over Lie groups), without the need for capsules. Previous work on group convolutions h

as been hampered by strong assumptions about the group, which precludes the appl ication of such techniques to common warps in computer vision such as affine and homographic. Our framework enables the implementation of group convolutions ove r any finite-dimensional Lie group. We empirically validate our approach on the benchmark affine-invariant classification task, where we achieve 30% improvemen t in accuracy against conventional CNNs while outperforming most CapsNets. As fu rther illustration of the generality of our framework, we train a homography-con volutional model which achieves superior robustness on a homography-perturbed da taset, where CapsNet results degrade.
****

Unbiased Teacher v2: Semi-Supervised Object Detection for Anchor-Free and Anchor -Based Detectors
Yen-Cheng Liu, Chih-Yao Ma, Zsolt Kira; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9819-9828
With the recent development of Semi-Supervised Object Detection (SS-OD) techniqu es, object detectors can be improved by using a limited amount of labeled data a nd abundant unlabeled data. However, there are still two challenges that are not addressed: (1) there is no prior SS-OD work on anchor-free detectors, and (2) p rior works are ineffective when pseudo-labeling bounding box regression. In this paper, we present Unbiased Teacher v2, which shows the generalization of SS-OD method to anchor-free detectors and also introduces Listen2Student mechanism for the unsupervised regression loss. Specifically, we first present a study examin ing the effectiveness of existing SS-OD methods on anchor-free detectors and fin d that they achieve much lower performance improvements under the semi-supervise d setting. We also observe that box selection with centerness and the localizati on-based labeling used in anchor-free detectors cannot work well under the semi- supervised setting. On the other hand, our Listen2Student mechanism explicitly p revents misleading pseudo-labels in the training of bounding box regression; we specifically develop a novel pseudo-labeling selection mechanism based on the Te acher and Student's relative uncertainties. This idea contributes to favorable i mprovement in the regression branch in the semi-supervised setting. Our method, which works for both anchor-free and anchor-based methods, consistently performs favorably against the state-of-the-art methods in VOC, COCO-standard, and COCO- additional.
****

GPU-Based Homotopy Continuation for Minimal Problems in Computer Vision
Chiang-Heng Chien, Hongyi Fan, Ahmad Abdelfattah, Elias Tsigaridas, Stanimire To mov, Benjamin Kimia; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2022, pp. 15765-15776
Systems of polynomial equations arise frequently in computer vision, especially in multiview geometry problems. Traditional methods for solving these systems ty pically aim to eliminate variables to reach a univariate polynomial, e.g., a tenth-order polynomial for 5-point pose estimation, using clever manipulations, or more generally using Grobner basis, resultants, and elimination templates, le ading to successful algorithms for multiview geometry and other problems. Howeve r, these methods do not work when the problem is complex and when they do, they face efficiency and stability issues. Homotopy Continuation (HC) can solve more complex problems without the stability issues, and with guarantees of a global s olution, but they are known to be slow. In this paper we show that HC can be par allelized on a GPU, showing significant speedups up to 56 times on polynomial be nchmarks. We also show that GPU-HC can be generically applied to a range of comp uter vision problems, including 4-view triangulation and trifocal pose estimatio n with unknown focal length, which cannot be solved with elimination template bu t they can be efficiently solved with HC. GPU-HC opens the door to easy formulat ion and solution of a range of computer vision problems.
****

Learning Pixel-Level Distinctions for Video Highlight Detection
Fanyue Wei, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, Lixin Duan; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3073-3082

The goal of video highlight detection is to select the most attractive segments from a long video to depict the most interesting parts of the video. Existing methods typically focus on modeling relationship between different video segments in order to learning a model that can assign highlight scores to these segments; however, these approaches do not explicitly consider the contextual dependency within individual segments. To this end, we propose to learn pixel-level distinctions to improve the video highlight detection. This pixel-level distinction indicates whether or not each pixel in one video belongs to an interesting section. The advantages of modeling such fine-level distinctions are two-fold. First, it allows us to exploit the temporal and spatial relations of the content in one video, since the distinction of a pixel in one frame is highly dependent on both the content before this frame and the content around this pixel in this frame. Second, learning the pixel-level distinction also gives a good explanation to the video highlight task regarding what contents in a highlight segment will be attractive to people. We design an encoder-decoder network to estimate the pixel-level distinction, in which we leverage the 3D convolutional neural networks to exploit the temporal context information, and further take advantage of the visual saliency to model the spatial distinction. State-of-the-art performance on three public benchmarks clearly validates the effectiveness of our framework for video highlight detection.

****

Noise Distribution Adaptive Self-Supervised Image Denoising Using Tweedie Distribution and Score Matching

Kwanyoung Kim, Taesung Kwon, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2008-2016

Tweedie distributions are a special case of exponential dispersion models, which are often used in classical statistics as distributions for generalized linear models. Here, we reveal that Tweedie distributions also play key roles in modern deep learning era, leading to a distribution independent self-supervised image denoising formula without clean reference images. Specifically, by combining with the recent Noise2Score self-supervised image denoising approach and the saddle point approximation of Tweedie distribution, we can provide a general closed-form denoising formula that can be used for large classes of noise distributions without ever knowing the underlying noise distribution. Similar to the original Noise2Score, the new approach is composed of two successive steps: score matching using perturbed noisy images, followed by a closed form image denoising formula via distribution-independent Tweedie's formula. This also suggests a systematic algorithm to estimate the noise model and noise parameters for a given noisy image data set. Through extensive experiments, we demonstrate that the proposed method can accurately estimate noise models and parameters, and provide the state-of-the-art self-supervised image denoising performance in the benchmark dataset and real-world dataset.

****

Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation

Yihan Wang, Muyang Li, Han Cai, Wei-Ming Chen, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13126-13136

Pose estimation plays a critical role in human-centered vision applications. However, it is difficult to deploy state-of-the-art HRNet-based pose estimation models on resource-constrained edge devices due to the high computational cost (more than 150 GMACs per frame). In this paper, we study efficient architecture design for real-time multi-person pose estimation on edge. We reveal that HRNet's high-resolution branches are redundant for models at the low-computation region via our gradual shrinking experiments. Removing them improves both efficiency and performance. Inspired by this finding, we design LitePose, an efficient single-branch architecture for pose estimation, and introduce two simple approaches to enhance the capacity of LitePose, including fusion deconv head and large kernel conv. On mobile platforms, LitePose reduces the latency by up to 5.0x without sacrificing performance, compared with prior state-of-the-art efficient pose estimation models, pushing the frontier of real-time multi-person pose estimation on e

dge. Our code and pre-trained models are released at https://github.com/mit-han-lab/litepose.
****

Boosting Black-Box Attack With Partially Transferred Conditional Adversarial Distribution
Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, Shu-Tao Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15095-15104
This work studies black-box adversarial attacks against deep neural networks (DNNs), where the attacker can only access the query feedback returned by the attacked DNN model, while other information such as model parameters or the training datasets are unknown. One promising approach to improve attack performance is utilizing the adversarial transferability between some white-box surrogate models and the target model (i.e., the attacked model). However, due to the possible differences on model architectures and training datasets between surrogate and target models, dubbed "surrogate biases", the contribution of adversarial transferability to improving the attack performance may be weakened. To tackle this issue, we innovatively propose a black-box attack method by developing a novel mechanism of adversarial transferability, which is robust to the surrogate biases. The general idea is transferring partial parameters of the conditional adversarial distribution (CAD) of surrogate models, while learning the untransferred parameters based on queries to the target model, to keep the flexibility to adjust the CAD of the target model on any new benign sample. Extensive experiments on benchmark datasets and attacking against real-world API demonstrate the superior attack performance of the proposed method.
****

CLIPstyler: Image Style Transfer With a Single Text Condition
Gihyun Kwon, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18062-18071
Existing neural style transfer methods require reference style images to transfer texture information of style images to content images. However, in many practical situations, users may not have reference style images but still be interested in transferring styles by just imagining them. In order to deal with such applications, we propose a new framework that enables a style transfer 'without' a style image, but only with a text description of the desired style. Using the pre-trained text-image embedding model of CLIP, we demonstrate the modulation of the style of content images only with a single text condition. Specifically, we propose a patch-wise text-image matching loss with multiview augmentations for realistic texture transfer. Extensive experimental results confirmed the successful image style transfer with realistic textures that reflect semantic query texts.
****

Ray Priors Through Reprojection: Improving Neural Radiance Fields for Novel View Extrapolation
Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, Xing Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18376-18386
Neural Radiance Fields (NeRF) have emerged as a potent paradigm for representing scenes and synthesizing photo-realistic images. A main limitation of conventional NeRFs is that they often fail to produce high-quality renderings under novel viewpoints that are significantly different from the training viewpoints. In this paper, instead of exploiting few-shot image synthesis, we study the novel view extrapolation setting that (1) the training images can well describe an object, and (2) there is a notable discrepancy between the training and test viewpoints' distributions. We present RapNeRF (RAy Priors) as a solution. Our insight is that the inherent appearances of a 3D surface's arbitrary visible projections should be consistent. We thus propose a random ray casting policy that allows training unseen views using seen views. Furthermore, we show that a ray atlas pre-computed from the observed rays' viewing directions could further enhance the rendering quality for extrapolated views. A main limitation is that RapNeRF would remove the strong view-dependent effects because it leverages the multi-view consis

tency property.
****
Spatio-Temporal Relation Modeling for Few-Shot Action Recognition
Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19958-19967
We propose a novel few-shot action recognition framework, STRM, which enhances class-specific feature discriminability while simultaneously learning higher-order temporal representations. The focus of our approach is a novel spatio-temporal enrichment module that aggregates spatial and temporal contexts with dedicated local patch-level and global frame-level feature enrichment sub-modules. Local patch-level enrichment captures the appearance-based characteristics of actions. On the other hand, global frame-level enrichment explicitly encodes the broad temporal context, thereby capturing the relevant object features over time. The resulting spatio-temporally enriched representations are then utilized to learn the relational matching between query and support action sub-sequences. We further introduce a query-class similarity classifier on the patch-level enriched features to enhance class-specific feature discriminability by reinforcing the feature learning at different stages in the proposed framework. Experiments are performed on four few-shot action recognition benchmarks: Kinetics, SSv2, HMDB51 and UCF101. Our extensive ablation study reveals the benefits of the proposed contributions. Furthermore, our approach sets a new state-of-the-art on all four benchmarks. On the challenging SSv2 benchmark, our approach achieves an absolute gain of 3.5% in classification accuracy, as compared to the best existing method in the literature. Our code and models are available at https://github.com/Anirudh257/strm.
****
Pop-Out Motion: 3D-Aware Image Deformation via Learning the Shape Laplacian
Jihyun Lee, Minhyuk Sung, Hyunjin Kim, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18532-18541
We propose a framework that can deform an object in a 2D image as it exists in 3D space. Most existing methods for 3D-aware image manipulation are limited to (1) only changing the global scene information or depth, or (2) manipulating an object of specific categories. In this paper, we present a 3D-aware image deformation method with minimal restrictions on shape category and deformation type. While our framework leverages 2D-to-3D reconstruction, we argue that reconstruction is not sufficient for realistic deformations due to the vulnerability to topological errors. Thus, we propose to take a supervised learning-based approach to predict the shape Laplacian of the underlying volume of a 3D reconstruction represented as a point cloud. Given the deformation energy calculated using the predicted shape Laplacian and user-defined deformation handles (e.g., keypoints), we obtain bounded biharmonic weights to model plausible handle-based image deformation. In the experiments, we present our results of deforming 2D character and clothed human images. We also quanti- tatively show that our approach can produce more accurate deformation weights compared to alternative methods (i.e., mesh reconstruction and point cloud Laplacian methods).
****
Volumetric Bundle Adjustment for Online Photorealistic Scene Capture
Ronald Clark; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6124-6132
Efficient photorealistic scene capture is a challenging task. Current online reconstruction systems can operate very efficiently, but images generated from the models captured by these systems are often not photorealistic. Recent approaches based on neural volume rendering can render novel views at high fidelity, but they often require a long time to train, making them impractical for applications that require real-time scene capture. In this paper, we propose a system that can reconstruct photorealistic models of complex scenes in an efficient manner. Our system processes images online, i.e. it can obtain a good quality estimate of both the scene geometry and appearance at roughly the same rate the video is ca

ptured. To achieve the efficiency, we propose a hierarchical feature volume usin g VDB grids. This representation is memory efficient and allows for fast queryin g of the scene information. Secondly, we introduce a novel optimization techniqu e that improves the efficiency of the bundle adjustment which allows our system to converge to the target camera poses and scene geometry much faster. Experimen ts on real-world scenes show that our method outperforms existing systems in ter ms of efficiency and capture quality. To the best of our knowledge, this is the first method that can achieve online photorealistic scene capture.
****

Multi-Person Extreme Motion Prediction
Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 13053-13064
Human motion prediction aims to forecast future poses given a sequence of past 3 D skeletons. While this problem has recently received increasing attention, it h as mostly been tackled for single humans in isolation. In this paper, we explore this problem when dealing with humans performing collaborative tasks, we seek t o predict the future motion of two interacted persons given two sequences of the ir past skeletons. We propose a novel cross interaction attention mechanism that exploits historical information of both persons, and learns to predict cross de pendencies between the two pose sequences. Since no dataset to train such intera ctive situations is available, we collected ExPI (Extreme Pose Interaction) data set, a new lab-based person interaction dataset of professional dancers performi ng Lindy-hop dancing actions, which contains 115 sequences with 30K frames annot ated with 3D body poses and shapes. We thoroughly evaluate our cross interaction network on ExPI and show that both in short- and long-term predictions, it cons istently outperforms state-of-the-art methods for single-person motion predictio n. Our code and dataset are available at: https://team.inria.fr/robotlearn/multi -person-extreme-motion-prediction/.
****

Masking Adversarial Damage: Finding Adversarial Saliency for Robust and Sparse N etwork
Byung-Kwan Lee, Junho Kim, Yong Man Ro; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15126-15136
Adversarial examples provoke weak reliability and potential security issues in d eep neural networks. Although adversarial training has been widely studied to im prove adversarial robustness, it works in an over-parameterized regime and requi res high computations and large memory budgets. To bridge adversarial robustness and model compression, we propose a novel adversarial pruning method, Masking A dversarial Damage (MAD) that employs second-order information of adversarial los s. By using it, we can accurately estimate adversarial saliency for model parame ters and determine which parameters can be pruned without weakening adversarial robustness. Furthermore, we reveal that model parameters of initial layer are hi ghly sensitive to the adversarial examples and show that compressed feature repr esentation retains semantic information for the target objects. Through extensiv e experiments on three public datasets, we demonstrate that MAD effectively prun es adversarially trained networks without loosing adversarial robustness and sho ws better performance than previous adversarial pruning methods.
****

Channel Balancing for Accurate Quantization of Winograd Convolutions
Vladimir Chikin, Vladimir Kryzhanovskiy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12507-12516
It is well known that Winograd convolution algorithms speed up the widely used s mall-size convolutions. However, the problem of quantization of Winograd convolu tions is challenging - while quantization of slower Winograd algorithms does not cause problems, quantization of faster Winograd algorithms often leads to a sig nificant drop in the quality of models. We introduce a novel class of Winograd a lgorithms that balances the filter and input channels in the Winograd domain. Un like traditional Winograd convolutions, the proposed convolution balances the ra nges of input channels on the forward pass by scaling the input tensor using spe

cial balancing coefficients (the filter channels are balanced offline). As a res ult of balancing, the inputs and filters of the Winograd convolution are much ea sier to quantize. Thus, the proposed technique allows us to obtain models with q uantized Winograd convolutions, the quality of which is significantly higher tha n the quality of models with traditional quantized Winograd convolutions. Moreov er, we propose a special direct algorithm for calculating the balancing coeffici ents, which does not require additional model training. This algorithm makes it easy to obtain the post-training quantized balanced Winograd convolutions - one should just feed a few data samples to the model without training to calibrate s pecial parameters. In addition, it is possible to initialize the balancing coeff icients using this algorithm and further train them as trainable variables durin g Winograd quantization-aware training for greater quality improvement.
****

RegNeRF: Regularizing Neural Radiance Fields for View Synthesis From Sparse Inpu ts
Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andre as Geiger, Noha Radwan; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 5480-5490
Neural Radiance Fields (NeRF) have emerged as a powerful representation for the task of novel view synthesis due to their simplicity and state-of-the-art perfor mance. Though NeRF can produce photorealistic renderings of unseen viewpoints wh en many input views are available, its performance drops significantly when this number is reduced. We observe that the majority of artifacts in sparse input sc enarios are caused by errors in the estimated scene geometry, and by divergent b ehavior at the start of training. We address this by regularizing the geometry a nd appearance of patches rendered from unobserved viewpoints, and annealing the ray sampling space during training. We additionally use a normalizing flow model to regularize the color of unobserved viewpoints. Our model outperforms not onl y other methods that optimize over a single scene, but in many cases also condit ional models that are extensively pre-trained on large multi-view datasets.
****

Structured Local Radiance Fields for Human Avatar Modeling
Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, Yebin Liu; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2022, pp. 15893-15903
It is extremely challenging to create an animatable clothed human avatar from RG B videos, especially for loose clothes due to the difficulties in motion modelin g. To address this problem, we introduce a novel representation on the basis of recent neural scene rendering techniques. The core of our representation is a se t of structured local radiance fields, which are anchored to the pre-defined nod es sampled on a statistical human body template. These local radiance fields not only leverage the flexibility of implicit representation in shape and appearanc e modeling, but also factorize cloth deformations into skeleton motions, node re sidual translations and the dynamic detail variations inside each individual rad iance field. To learn our representation from RGB data and facilitate pose gener alization, we propose to learn the node translations and the detail variations i n a conditional generative latent space. Overall, our method enables automatic c onstruction of animatable human avatars for various types of clothes without the need for scanning subject-specific templates, and can generate realistic images with dynamic details for novel poses. Experiment show that our method outperfor ms state-of-the-art methods both qualitatively and quantitatively.
****

Towards Noiseless Object Contours for Weakly Supervised Semantic Segmentation
Jing Li, Junsong Fan, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16856-16865
Image-level label based weakly supervised semantic segmentation has attracted mu ch attention since image labels are very easy to obtain. Existing methods usuall y generate pseudo labels from class activation map (CAM) and then train a segmen tation model. CAM usually highlights partial objects and produce incomplete pseu do labels. Some methods explore object contour by training a contour model with

CAM seed label supervision and then propagate CAM score from discriminative regi
ons to non-discriminative regions with contour guidance. The propagation process
 suffers from the noisy intra-object contours, and inadequate propagation result
s produce incomplete pseudo labels. This is because the coarse CAM seed label la
cks sufficient precise semantic information to suppress contour noise. In this p
aper, we train a SANCE model which utilizes an auxiliary segmentation module to
supplement high-level semantic information for contour training by backbone feat
ure sharing and online label supervision. The auxiliary segmentation module also
 provides more accurate localization map than CAM for pseudo label generation. W
e evaluate our approach on Pascal VOC 2012 and MS COCO 2014 benchmarks and achie
ve state-of-the-art performance, demonstrating the effectiveness of our method.
****

Ranking-Based Siamese Visual Tracking
Feng Tang, Qiang Ling; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2022, pp. 8741-8750
Current Siamese-based trackers mainly formulate the visual tracking into two ind
epedent subtasks, including classification and localization. They learn the clas
sification subnetwork by processing each sample separately and neglect the relat
ionship among positive and negative samples. Moreover, such tracking paradigm ta
kes only the classification confidence of proposals for the final prediction, wh
ich may yield the misalignment between classification and localization. To resol
ve these issues, this paper proposes a ranking-based optimization algorithm to e
xplore the relationship among different proposals. To this end, we introduce two
 ranking losses, including the classification one and the IoU-guided one, as opt
imization constraints. The classification ranking loss can ensure that positive
samples rank higher than hard negative ones, i.e., distractors, so that the trac
kers can select the foreground samples successfully without being fooled by the
distractors. The IoU-guided ranking loss aims to align classification confidence
 scores with the Intersection over Union(IoU) of the corresponding localization
prediction for positive samples, enabling the well-localized prediction to be re
presented by high classification confidence. Specifically, the proposed two rank
ing losses are compatible with most Siamese trackers and incur no additional com
putation for inference. Extensive experiments on seven tracking benchmarks, incl
uding OTB100, UAV123, TC128, VOT2016, NFS30, GOT-10k and LaSOT, demonstrate the
effectiveness of the proposed ranking-based optimization algorithm.
****

Learnable Lookup Table for Neural Network Quantization
Longguang Wang, Xiaoyu Dong, Yingqian Wang, Li Liu, Wei An, Yulan Guo; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 12423-12433
Neural network quantization aims at reducing bit-widths of weights and activatio
ns for memory and computational efficiency. Since a linear quantizer (i.e., roun
d(*) function) cannot well fit the bell-shaped distributions of weights and acti
vations, many existing methods use pre-defined functions (e.g., exponential func
tion) with learnable parameters to build the quantizer for joint optimization. H
owever, these complicated quantizers introduce considerable computational overhe
ad during inference since activation quantization should be conducted online. In
 this paper, we formulate the quantization process as a simple lookup operation
and propose to learn lookup tables as quantizers. Specifically, we develop diffe
rentiable lookup tables and introduce several training strategies for optimizati
on. Our lookup tables can be trained with the network in an end-to-end manner to
 fit the distributions in different layers and have very small additional comput
ational cost. Comparison with previous methods show that quantized networks usin
g our lookup tables achieve state-of-the-art performance on image classification
, image super-resolution, and point cloud classification tasks.
****

SEEG: Semantic Energized Co-Speech Gesture Generation
Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, Yi Yang; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 10473-10482

Talking gesture generation is a practical yet challenging task which aims to synthesize gestures in line with speech. Gestures with meaningful signs can better convey useful information and arouse sympathy in the audience. Current works focus on aligning gestures with the speech rhythms, which are hard to mine the semantics and model semantic gestures explicitly. In this paper, we propose a novel method SEmantic Energized Generation (SEEG), for semantic-aware gesture generation. Our method contains two parts: DEcoupled Mining module (DEM) and Semantic Energizing Module (SEM). DEM decouples the semantic-irrelevant information from inputs and separately mines information for the beat and semantic gestures. SEM conducts semantic learning and produces semantic gestures. Apart from representational similarity, SEM requires the predictions to express the same semantics as the ground truth. Besides, a semantic prompter is designed in SEM to leverage the semantic-aware supervision to predictions. This promotes the networks to learn and generate semantic gestures. Experimental results reported in three metrics on different benchmarks prove that SEEG efficiently mines semantic cues and generates semantic gestures. In comparison, SEEG outperforms other methods in all semantic-aware evaluations on different datasets. Qualitative evaluations also indicate the superiority of SEEG in semantic expressiveness.

****

AdaViT: Adaptive Vision Transformers for Efficient Image Recognition

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12309-12318

Built on top of self-attention mechanisms, vision transformers have demonstrated remarkable performance on a variety of vision tasks recently. While achieving excellent performance, they still require relatively intensive computational cost that scales up drastically as the numbers of patches, self-attention heads and transformer blocks increase. In this paper, we argue that due to the large variations among images, their need for modeling long-range dependencies between patches differ. To this end, we introduce AdaViT, an adaptive computation framework that learns to derive usage policies on which patches, self-attention heads and transformer blocks to use throughout the backbone on a per-input basis, aiming to improve inference efficiency of vision transformers with a minimal drop of accuracy for image recognition. Optimized jointly with a transformer backbone in an end-to-end manner, a light-weight decision network is attached to the backbone to produce decisions on-the-fly. Extensive experiments on ImageNet demonstrate that our method obtains more than 2x improvement on efficiency compared to state-of-the-art vision transformers with only 0.8% drop of accuracy, achieving good efficiency/accuracy trade-offs conditioned on different computational budgets. We further conduct quantitative and qualitative analysis on learned usage polices and provide more insights on the redundancy in vision transformers.

****

Compound Domain Generalization via Meta-Knowledge Encoding

Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7119-7129

Domain generalization (DG) aims to improve the generalization performance for an unseen target domain by using the knowledge of multiple seen source domains. Mainstream DG methods typically assume that the domain label of each source sample is known a priori, which is challenged to be satisfied in many real-world applications. In this paper, we study a practical problem of compound DG, which relaxes the discrete domain assumption to the mixed source domains setting. On the other hand, current DG algorithms prioritize the focus on semantic invariance across domains (one-vs-one), while paying less attention to the holistic semantic structure (many-vs-many). Such holistic semantic structure, referred to as meta-knowledge here, is crucial for learning generalizable representations. To this end, we present Compound Domain Generalization via Meta-Knowledge Encoding (COMEN), a general approach to automatically discover and model latent domains in two steps. Firstly, we introduce Style-induced Domain-specific Normalization (SDNorm) to re-normalize the multi-modal underlying distributions, thereby dividing the m

ixture of source domains into latent clusters. Secondly, we harness the prototyp
e representations, the centroids of classes, to perform relational modeling in t
he embedding space with two parallel and complementary modules, which explicitly
 encode the semantic structure for the out-of-distribution generalization. Exper
iments on four standard DG benchmarks reveal that COMEN exceeds the state-of-the
-art performance without the need of domain supervision.
****

NAN: Noise-Aware NeRFs for Burst-Denoising
Naama Pearl, Tali Treibitz, Simon Korman; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12672-12681
Burst denoising is now more relevant than ever, as computational photography hel
ps overcome sensitivity issues inherent in mobile phones and small cameras. A ma
jor challenge in burst-denoising is in coping with pixel misalignment, which was
 so far handled with rather simplistic assumptions of simple motion, or the abil
ity to align in pre-processing. Such assumptions are not realistic in the presen
ce of large motion and high levels of noise. We show that Neural Radiance Fields
 (NeRFs), originally suggested for physics-based novel-view rendering, can serve
 as a powerful framework for burst denoising. NeRFs have an inherent capability
of handling noise as they integrate information from multiple images, but they a
re limited in doing so, mainly since they build on pixel-wise operations which a
re suitable to ideal imaging conditions. Our approach, termed NAN, leverages int
er-view and spatial information in NeRFs to better deal with noise. It achieves
state-of-the-art results in burst denoising and is especially successful in copi
ng with large movement and occlusions, under very high levels of noise. With the
 rapid advances in accelerating NeRFs, it could provide a powerful platform for
denoising in challenging environments.
****

Physical Inertial Poser (PIP): Physics-Aware Real-Time Human Motion Tracking Fro
m Sparse Inertial Sensors
Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christ
ian Theobalt, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2022, pp. 13167-13178
Motion capture from sparse inertial sensors has shown great potential compared t
o image-based approaches since occlusions do not lead to a reduced tracking qual
ity and the recording space is not restricted to be within the viewing frustum o
f the camera. However, capturing the motion and global position only from a spar
se set of inertial sensors is inherently ambiguous and challenging. In consequen
ce, recent state-of-the-art methods can barely handle very long period motions,
and unrealistic artifacts are common due to the unawareness of physical constrai
nts. To this end, we present the first method which combines a neural kinematics
 estimator and a physics-aware motion optimizer to track body motions with only
6 inertial sensors. The kinematics module first regresses the motion status as a
 reference, and then the physics module refines the motion to satisfy the physic
al constraints. Experiments demonstrate a clear improvement over the state of th
e art in terms of capture accuracy, temporal stability, and physical correctness
.
****

b-DARTS: Beta-Decay Regularization for Differentiable Architecture Search
Peng Ye, Baopu Li, Yikang Li, Tao Chen, Jiayuan Fan, Wanli Ouyang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
2, pp. 10874-10883
Neural Architecture Search (NAS) has attracted increasingly more attention in re
cent years because of its capability to design deep neural network automatically
. Among them, differential NAS approaches such as DARTS, have gained popularity
for the search efficiency. However, they suffer from two main issues, the weak r
obustness to the performance collapse and the poor generalization ability of the
 searched architectures. To solve these two problems, a simple-but-efficient reg
ularization method, termed as Beta-Decay, is proposed to regularize the DARTS-ba
sed NAS searching process. Specifically, Beta-Decay regularization can impose co
nstraints to keep the value and variance of activated architecture parameters fr

om too large. Furthermore, we provide in-depth theoretical analysis on how it wo rks and why it works. Experimental results on NAS-Bench-201 show that our propos ed method can help to stabilize the searching process and makes the searched net work more transferable across different datasets. In addition, our search scheme shows an outstanding property of being less dependent on training time and data . Comprehensive experiments on a variety of search spaces and datasets validate the effectiveness of the proposed method. The code is available at https://githu b.com/Sunshine-Ye/Beta-DARTS.

****

Vector Quantized Diffusion Model for Text-to-Image Synthesis

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 10696-10706

We present the vector quantized diffusion (VQ-Diffusion) model for text-to-image generation. This method is based on a vector quantized variational autoencoder (VQ-VAE) whose latent space is modeled by a conditional variant of the recently developed Denoising Diffusion Probabilistic Model (DDPM). We find that this late nt-space method is well-suited for text-to-image generation tasks because it not only eliminates the unidirectional bias with existing methods but also allows u s to incorporate a mask-and-replace diffusion strategy to avoid the accumulation of errors, which is a serious problem with existing methods. Our experiments sh ow that the VQ-Diffusion produces significantly better text-to-image generation results when compared with conventional autoregressive (AR) models with similar numbers of parameters. Compared with previous GAN-based text-to-image methods, o ur VQ-Diffusion can handle more complex scenes and improve the synthesized image quality by a large margin. Finally, we show that the image generation computati on in our method can be made highly efficient by reparameterization. With tradit ional AR methods, the text-to-image generation time increases linearly with the output image resolution and hence is quite time consuming even for normal size i mages. The VQ-Diffusion allows us to achieve a better trade-off between quality and speed. Our experiments indicate that the VQ-Diffusion model with the reparam eterization is fifteen times faster than traditional AR methods while achieving a better image quality.

****

CMT: Convolutional Neural Networks Meet Vision Transformers

Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, Chang Xu; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12175-12185

Vision transformers have been successfully applied to image recognition tasks du e to their ability to capture long-range dependencies within an image. However, there are still gaps in both performance and computational cost between transfor mers and existing convolutional neural networks (CNNs). In this paper, we aim to address this issue and develop a network that can outperform not only the canon ical transformers, but also the high-performance convolutional models. We propos e a new transformer based hybrid network by taking advantage of transformers to capture long-range dependencies, and of CNNs to extract local information. Furth ermore, we scale it to obtain a family of models, called CMTs, obtaining much be tter trade-off for accuracy and efficiency than previous CNN-based and transform er-based models. In particular, our CMT-S achieves 83.5% top-1 accuracy on Image Net, while being 14x and 2x smaller on FLOPs than the existing DeiT and Efficien tNet, respectively. The proposed CMT-S also generalizes well on CIFAR10 (99.2%), CIFAR100 (91.7%), Flowers (98.7%), and other challenging vision datasets such a s COCO (44.3% mAP), with considerably less computational cost. Code is available at https://github.com/ggjy/CMT.pytorch.

****

Hyperspherical Consistency Regularization

Cheng Tan, Zhangyang Gao, Lirong Wu, Siyuan Li, Stan Z. Li; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7244-7255

Recent advances in contrastive learning have enlightened diverse applications ac

ross various semi-supervised fields. Jointly training supervised learning and un supervised learning with a shared feature encoder becomes a common scheme. Thoug h it benefits from taking advantage of both feature-dependent information from s elf-supervised learning and label-dependent information from supervised learning , this scheme remains suffering from bias of the classifier. In this work, we sy stematically explore the relationship between self-supervised learning and super vised learning, and study how self-supervised learning helps robust data-efficie nt deep learning. We propose hyperspherical consistency regularization (HCR), a simple yet effective plug-and-play method, to regularize the classifier using fe ature-dependent information and thus avoid bias from labels. Specifically, HCR f irst project logits from the classifier and feature projections from the project ion head on the respective hypersphere, then it enforces data points on hypersph eres to have similar structures by minimizing binary cross entropy of pairwise d istances' similarity metrics. Extensive experiments on semi-supervised learning and weakly-supervised learning demonstrate the effectiveness of our proposed met hod, by showing superior performance with HCR.

****

Unsupervised Image-to-Image Translation With Generative Prior

Shuai Yang, Liming Jiang, Ziwei Liu, Chen Change Loy; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18332-18341

Unsupervised image-to-image translation aims to learn the translation between tw o visual domains without paired data. Despite the recent progress in image trans lation models, it remains challenging to build mappings between complex domains with drastic visual discrepancies. In this work, we present a novel framework, G enerative Prior-guided UNsupervised Image-to-image Translation (GP-UNIT), to imp rove the overall quality and applicability of the translation algorithm. Our key insight is to leverage the generative prior from pre-trained class-conditional GANs (e.g., BigGAN) to learn rich content correspondences across various domains . We propose a novel coarse-to-fine scheme: we first distill the generative prio r to capture a robust coarse-level content representation that can link objects at an abstract semantic level, based on which fine-level content features are ad aptively learned for more accurate multi-level content correspondences. Extensiv e experiments demonstrate the superiority of our versatile framework over state-of-the-art methods in robust, high-quality and diversified translations, even fo r challenging and distant domains.

****

KNN Local Attention for Image Restoration

Hunsang Lee, Hyesong Choi, Kwanghoon Sohn, Dongbo Min; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2139-2149

Recent works attempt to integrate the non-local operation with CNNs or Transform er, achieving remarkable performance in image restoration tasks. The global simi larity, however, has the problems of the lack of locality and the high computati onal complexity that is quadratic to an input resolution. The local attention me chanism alleviates these issues by introducing the inductive bias of the localit y with convolution-like operators. However, by focusing only on adjacent positio ns, the local attention suffers from an insufficient receptive field for image r estoration. In this paper, we propose a new attention mechanism for image restor ation, called k-NN Image Transformer (KiT), that rectifies above mentioned limit ations. Specifically, the KiT groups k-nearest neighbor patches with locality se nsitive hashing (LSH), and the grouped patches are aggregated into each query pa tch by performing a pair-wise local attention. In this way, the pair-wise operat ion establishes non-local connectivity while maintaining the desired properties of the local attention, i.e., inductive bias of locality and linear complexity t o input resolution. The proposed method outperforms state-of-the-art restoration approaches on image denoising, deblurring and deraining benchmarks. The code wi ll be available at https://sites.google.com/view/cvpr22-kit.

****

Face Relighting With Geometrically Consistent Shadows

Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, Xiaoming Liu; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 4217-4226
Most face relighting methods are able to handle diffuse shadows, but struggle to handle hard shadows, such as those cast by the nose. Methods that propose techn iques for handling hard shadows often do not produce geometrically consistent sh adows since they do not directly leverage the estimated face geometry while synt hesizing them. We propose a novel differentiable algorithm for synthesizing hard shadows based on ray tracing, which we incorporate into training our face relig hting model. Our proposed algorithm directly utilizes the estimated face geometr y to synthesize geometrically consistent hard shadows. We demonstrate through qu antitative and qualitative experiments on Multi-PIE and FFHQ that our method pro duces more geometrically consistent shadows than previous face relighting method s while also achieving state-of-the-art face relighting performance under direct ional lighting. In addition, we demonstrate that our differentiable hard shadow modeling improves the quality of the estimated face geometry over diffuse shadin g models.
****

Open-Set Text Recognition via Character-Context Decoupling
Chang Liu, Chun Yang, Xu-Cheng Yin; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2022, pp. 4523-4532
The open-set text recognition task is an emerging challenge that requires an ext ra capability to cognize novel characters during evaluation. We argue that a maj or cause of the limited performance for current methods is the confounding effec t of contextual information over the visual information of individual characters . Under open-set scenarios, the intractable bias in contextual information can b e passed down to visual information, consequently impairing the classification p erformance. In this paper, a Character-Context Decoupling framework is proposed to alleviate this problem by separating contextual information and character-vis ual information. Contextual information can be decomposed into temporal informat ion and linguistic information. Here, temporal information that models character order and word length is isolated with a detached temporal attention module. Li nguistic information that models n-gram and other linguistic statistics is separ ated with a decoupled context anchor mechanism. A variety of quantitative and qu alitative experiments show that our method achieves promising performance on ope n-set, zero-shot, and close-set text recognition datasets.
****

Multi-Marginal Contrastive Learning for Multi-Label Subcellular Protein Localiza tion
Ziyi Liu, Zengmao Wang, Bo Du; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 20626-20635
Protein subcellular localization(PSL) is an important task to study human cell f unctions and cancer pathogenesis. It has attracted great attention in the comput er vision community. However, the huge size of immune histochemical (IHC) images , the disorganized location distribution in different tissue images and the limi ted training images are always the challenges for the PSL to learn a strong gene ralization model with deep learning. In this paper, we propose a deep protein su bcellular localization method with multi-marginal contrastive learning to percei ve the same PSLs in different tissue images and different PSLs within the same t issue image. In the proposed method, we learn the representation of an IHC image by fusing the global features from the downsampled images and local features fr om the selected patches with the activation map to tackle the oversize of an IHC image. Then a multi-marginal attention mechanism is proposed to generate contra stive pairs with different margins and improve the discriminative features of PS L patterns effectively. Finally, the ensemble prediction of each IHC image is ob tained with different patches. The results on the benchmark datasets show that t he proposed method achieves the significant improvements for the PSL task.
****

Probabilistic Warp Consistency for Weakly-Supervised Semantic Correspondences
Prune Truong, Martin Danelljan, Fisher Yu, Luc Van Gool; Proceedings of the IEEE

/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8708-8718
We propose Probabilistic Warp Consistency, a weakly-supervised learning objective for semantic matching. Our approach directly supervises the dense matching scores predicted by the network, encoded as a conditional probability distribution. We first construct an image triplet by applying a known warp to one of the images in a pair depicting different instances of the same object class. Our probabilistic learning objectives are then derived using the constraints arising from the resulting image triplet. We further account for occlusion and background clutter present in real image pairs by extending our probabilistic output space with a learnable unmatched state. To supervise it, we design an objective between image pairs depicting different object classes. We validate our method by applying it to four recent semantic matching architectures. Our weakly-supervised approach sets a new state-of-the-art on four challenging semantic matching benchmarks. Lastly, we demonstrate that our objective also brings substantial improvements in the strongly-supervised regime, when combined with keypoint annotations.
****

Predict, Prevent, and Evaluate: Disentangled Text-Driven Image Manipulation Empowered by Pre-Trained Vision-Language Model
Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18229-18238
To achieve disentangled image manipulation, previous works depend heavily on manual annotation. Meanwhile, the available manipulations are limited to a pre-defined set the models were trained for. We propose a novel framework, i.e., Predict, Prevent, and Evaluate (PPE), for disentangled text-driven image manipulation that requires little manual annotation while being applicable to a wide variety of manipulations. Our method approaches the targets by deeply exploiting the power of the large-scale pre-trained vision-language model CLIP. Concretely, we firstly Predict the possibly entangled attributes for a given text command. Then, based on the predicted attributes, we introduce an entanglement loss to Prevent entanglements during training. Finally, we propose a new evaluation metric to Evaluate the disentangled image manipulation. We verify the effectiveness of our method on the challenging face editing task. Extensive experiments show that the proposed PPE framework achieves much better quantitative and qualitative results than the up-to-date StyleCLIP baseline. Code is available at https://github.com/zipengxuc/PPE.
****

Optimizing Elimination Templates by Greedy Parameter Search
Evgeniy Martyushev, Jana Vráblíková, Tomas Pajdla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15754-15764
We propose a new method for constructing elimination templates for efficient polynomial system solving of minimal problems in structure from motion, image matching, and camera tracking. We first construct a particular affine parameterization of the elimination templates for systems with a finite number of distinct solutions. Then, we use a heuristic greedy optimization strategy over the space of parameters to get a template with a small size. We test our method on 34 minimal problems in computer vision. For all of them, we found the templates either of the same or smaller size compared to the state-of-the-art. For some difficult examples, our templates are, e.g., 2.1, 2.5, 3.8, 6.6 times smaller. For the problem of refractive absolute pose estimation with unknown focal length, we have found a template that is 20 times smaller. Our experiments on synthetic data also show that the new solvers are fast and numerically accurate. We also present a fast and numerically accurate solver for the problem of relative pose estimation with unknown common focal length and radial distortion.
****

TransMix: Attend To Mix for Vision Transformers
Jie-Neng Chen, Shuyang Sun, Ju He, Philip H.S. Torr, Alan Yuille, Song Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C

VPR), 2022, pp. 12135-12144
Mixup-based augmentation has been found to be effective for generalizing models during training, especially for Vision Transformers (ViTs) since they can easily overfit. However, previous mixup-based methods have an underlying prior knowledge that the linearly interpolated ratio of targets should be kept the same as the ratio proposed in input interpolation. This may lead to a strange phenomenon that sometimes there is no valid object in the mixed image due to the random process in augmentation but there is still response in the label space. To bridge such gap between the input and label spaces, we propose TransMix, which mixes labels based on the attention maps of Vision Transformers. The confidence of the label will be larger if the corresponding input image is weighted higher by the attention map. TransMix is embarrassingly simple and can be implemented in just a few lines of code without introducing any extra parameters and FLOPs to ViT-based models. Experimental results show that our method can consistently improve various ViT-based models at scales on ImageNet classification. After pre-trained with TransMix on ImageNet, the ViT-based models also demonstrate better transferability to semantic segmentation, object detection and instance segmentation. TransMix also exhibits to be more robust when evaluating on 4 different benchmarks. Code is publicly available at https://github.com/Beckschen/TransMix.
****

HOP: History-and-Order Aware Pre-Training for Vision-and-Language Navigation
Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15418-15427
Pre-training has been adopted in a few of recent works for Vision-and-Language Navigation (VLN). However, previous pre-training methods for VLN either lack the ability to predict future actions or ignore the trajectory contexts, which are essential for a greedy navigation process. In this work, to promote the learning of spatio-temporal visual-textual correspondence as well as the agent's capability of decision making, we propose a novel history-and-order aware pre-training paradigm (HOP) with VLN-specific objectives that exploit the past observations and support future action prediction. Specifically, in addition to the commonly used Masked Language Modeling (MLM) and Trajectory-Instruction Matching (TIM), we design two proxy tasks to model temporal order information: Trajectory Order Modeling (TOM) and Group Order Modeling (GOM). Moreover, our navigation action prediction is also enhanced by introducing the task of Action Prediction with History (APH), which takes into account the history visual perceptions. Extensive experimental results on four downstream VLN tasks (R2R, REVERIE, NDH, RxR) demonstrate the effectiveness of our proposed method compared against several state-of-the-art agents.
****

Inertia-Guided Flow Completion and Style Fusion for Video Inpainting
Kaidong Zhang, Jingjing Fu, Dong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5982-5991
Physical objects have inertia, which resists changes in the velocity and motion direction. Inspired by this, we introduce inertia prior that optical flow, which reflects object motion in a local temporal window, keeps unchanged in the adjacent preceding or subsequent frame. We propose a flow completion network to align and aggregate flow features from the consecutive flow sequences based on the inertia prior. The corrupted flows are completed under the supervision of customized losses on reconstruction, flow smoothness, and consistent ternary census transform. The completed flows with high fidelity give rise to significant improvement on the video inpainting quality. Nevertheless, the existing flow-guided cross-frame warping methods fail to consider the lightening and sharpness variation across video frames, which leads to spatial incoherence after warping from other frames. To alleviate such problem, we propose the Adaptive Style Fusion Network (ASFN), which utilizes the style information extracted from the valid regions to guide the gradient refinement in the warped regions. Moreover, we design a data simulation pipeline to reduce the training difficulty of ASFN. Extensive experiments show the superiority of our method against the state-of-the-art methods qu

antitatively and qualitatively. The project page is at https://github.com/hitachinsk/ISVI
****

RU-Net: Regularized Unrolling Network for Scene Graph Generation
Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19457-19466
Scene graph generation (SGG) aims to detect objects and predict the relationships between each pair of objects. Existing SGG methods usually suffer from several issues, including 1) ambiguous object representations, as graph neural network-based message passing (GMP) modules are typically sensitive to spurious inter-node correlations, and 2) low diversity in relationship predictions due to severe class imbalance and a large number of missing annotations. To address both problems, in this paper, we propose a regularized unrolling network (RU-Net). We first study the relation between GMP and graph Laplacian denoising (GLD) from the perspective of the unrolling technique, determining that GMP can be formulated as a solver for GLD. Based on this observation, we propose an unrolled message passing module and introduce an l_p-based graph regularization to suppress spurious connections between nodes. Second, we propose a group diversity enhancement module that promotes the prediction diversity of relationships via rank maximization. Systematic experiments demonstrate that RU-Net is effective under a variety of settings and metrics. Furthermore, RU-Net achieves new state-of-the-arts on three popular databases: VG, VRD, and OI.
****

Long-Tailed Visual Recognition via Gaussian Clouded Logit Adjustment
Mengke Li, Yiu-ming Cheung, Yang Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6929-6938
Long-tailed data is still a big challenge for deep neural networks, even though they have achieved great success on balanced data. We observe that vanilla training on long-tailed data with cross-entropy loss makes the instance-rich head classes severely squeeze the spatial distribution of the tail classes, which leads to difficulty in classifying tail class samples. Furthermore, the original cross-entropy loss can only propagate gradient short-lively because the gradient in softmax form rapidly approaches zero as the logit difference increases. This phenomenon is called softmax saturation. It is unfavorable for training on balanced data, but can be utilized to adjust the validity of the samples in long-tailed data, thereby solving the distorted embedding space of long-tailed problems. To this end, this paper proposes the Gaussian clouded logit adjustment by Gaussian perturbation of different class logits with varied amplitude. We define the amplitude of perturbation as cloud size and set relatively large cloud sizes to tail classes. The large cloud size can reduce the softmax saturation and thereby making tail class samples more active as well as enlarging the embedding space. To alleviate the bias in a classifier, we therefore propose the class-based effective number sampling strategy with classifier re-training. Extensive experiments on benchmark datasets validate the superior performance of the proposed method. Source code is available at: https://github.com/Keke921/GCLLoss.
****

Image Animation With Perturbed Masks
Yoav Shalev, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3647-3656
We present a novel approach for image-animation of a source image by a driving video, both depicting the same type of object. We do not assume the existence of pose models and our method is able to animate arbitrary objects without the knowledge of the object's structure. Furthermore, both, the driving video and the source image are only seen during test-time. Our method is based on a shared mask generator, which separates the foreground object from its background, and captures the object's general pose and shape. To control the source of the identity of the output frame, we employ perturbations to interrupt the unwanted identity information on the driver's mask. A mask-refinement module then replaces the identity of the driver with the identity of the source. Conditioned on the source ima

ge, the transformed mask is then decoded by a multi-scale generator that renders a realistic image, in which the content of the source frame is animated by the pose in the driving video. Due to the lack of fully supervised data, we train on the task of reconstructing frames from the same video the source image is taken from. Our method is shown to greatly outperform the state-of-the-art methods on multiple benchmarks. Our code and samples are available at https://github.com/itsyoavshalev/Image-Animation-with-Perturbed-Masks.
****

Exploring the Equivalence of Siamese Self-Supervised Learning via a Unified Gradient Framework
Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14431-14440
Self-supervised learning has shown its great potential to extract powerful visual representations without human annotations. Various works are proposed to deal with self-supervised learning from different perspectives: (1) contrastive learning methods (e.g., MoCo, SimCLR) utilize both positive and negative samples to guide the training direction; (2) asymmetric network methods (e.g., BYOL, SimSiam) get rid of negative samples via the introduction of a predictor network and the stop-gradient operation; (3) feature decorrelation methods (e.g., Barlow Twins, VICReg) instead aim to reduce the redundancy between feature dimensions. These methods appear to be quite different in the designed loss functions from various motivations. The final accuracy numbers also vary, where different networks and tricks are utilized in different works. In this work, we demonstrate that these methods can be unified into the same form. Instead of comparing their loss functions, we derive a unified formula through gradient analysis. Furthermore, we conduct fair and detailed experiments to compare their performances. It turns out that there is little gap between these methods, and the use of momentum encoder is the key factor to boost performance. From this unified framework, we propose UniGrad, a simple but effective gradient form for self-supervised learning. It does not require a memory bank or a predictor network, but can still achieve state-of-the-art performance and easily adopt other training strategies. Extensive experiments on linear evaluation and many downstream tasks also show its effectiveness. Code shall be released.
****

Point Density-Aware Voxels for LiDAR 3D Object Detection
Jordan S. K. Hu, Tianshu Kuai, Steven L. Waslander; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8469-8478
LiDAR has become one of the primary 3D object detection sensors in autonomous driving. However, LiDAR's diverging point pattern with increasing distance results in a non-uniform sampled point cloud ill-suited to discretized volumetric feature extraction. Current methods either rely on voxelized point clouds or use inefficient farthest point sampling to mitigate detrimental effects caused by density variation but largely ignore point density as a feature and its predictable relationship with distance from the LiDAR sensor. Our proposed solution, Point Density-Aware Voxel network (PDV), is an end-to-end two stage LiDAR 3D object detection architecture that is designed to account for these point density variations. PDV efficiently localizes voxel features from the 3D sparse convolution backbone through voxel point centroids. The spatially localized voxel features are then aggregated through a density-aware RoI grid pooling module using kernel density estimation (KDE) and self-attention with point density positional encoding. Finally, we exploit LiDAR's point density to distance relationship to refine our final bounding box confidences. PDV outperforms all state-of-the-art methods on the Waymo Open Dataset and achieves competitive results on the KITTI dataset.
****

Integrating Language Guidance Into Vision-Based Deep Metric Learning
Karsten Roth, Oriol Vinyals, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16177-16189
Deep Metric Learning (DML) proposes to learn metric spaces which encode semantic

similarities as embedding space distances. These spaces should be transferable to classes beyond those seen during training. Commonly, DML methods task networks to solve contrastive ranking tasks defined over binary class assignments. However, such approaches ignore higher-level semantic relations between the actual classes. This causes learned embedding spaces to encode incomplete semantic context and misrepresent the semantic relation between classes, impacting the generalizability of the learned metric space. To tackle this issue, we propose a language guidance objective for visual similarity learning. Leveraging language embeddings of expert- and pseudo-classnames, we contextualize and realign visual representation spaces corresponding to meaningful language semantics for better semantic consistency. Extensive experiments and ablations provide a strong motivation for our proposed approach and show language guidance offering significant, model-agnostic improvements for DML, achieving competitive and state-of-the-art results on all benchmarks. Code available at github.com/ExplainableML/LanguageGuidance_for_DML.

****

PartGlot: Learning Shape Part Segmentation From Language Reference Games

Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J. Guibas, Minhyuk Sung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16505-16514

We introduce PartGlot, a neural framework and associated architectures for learning semantic part segmentation of 3D shape geometry, based solely on part referential language. We exploit the fact that linguistic descriptions of a shape can provide priors on the shape's parts -- as natural language has evolved to reflect human perception of the compositional structure of objects, essential to their recognition and use. For training we use the paired geometry / language data collected in the ShapeGlot work for their reference game, where a speaker creates an utterance to differentiate a target shape from two distractors and the listener has to find the target based on this utterance. Our network is designed to solve this target discrimination problem, carefully incorporating a Transformer-based attention module so that the output attention can precisely highlight the semantic part or parts described in the language. Furthermore, the network operates without any direct supervision on the 3D geometry itself. Surprisingly, we further demonstrate that the learned part information is generalizable to shape classes unseen during training.Our approach opens the possibility of learning 3D shape parts from language alone, without the need for large-scale part geometry annotations, thus facilitating annotation acquisition.

****

Domain Generalization via Shuffled Style Assembly for Face Anti-Spoofing

Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, Zhongyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4123-4133

With diverse presentation attacks emerging continually, generalizable face anti-spoofing (FAS) has drawn growing attention. Most existing methods implement domain generalization (DG) on the complete representations. However, different image statistics may have unique properties for the FAS tasks. In this work, we separate the complete representation into content and style ones. A novel Shuffled Style Assembly Network (SSAN) is proposed to extract and reassemble different content and style features for a stylized feature space. Then, to obtain a generalized representation, a contrastive learning strategy is developed to emphasize liveness-related style information while suppress the domain-specific one. Finally, the representations of the correct assemblies are used to distinguish between living and spoofing during the inferring. On the other hand, despite the decent performance, there still exists a gap between academia and industry, due to the difference in data quantity and distribution. Thus, a new large-scale benchmark for FAS is built up to further evaluate the performance of algorithms in reality. Both qualitative and quantitative results on existing and proposed benchmarks demonstrate the effectiveness of our methods. The codes will be available at https://github.com/wangzhuo2019/SSAN.

****

A Simple Episodic Linear Probe Improves Visual Recognition in the Wild
Yuanzhi Liang, Linchao Zhu, Xiaohan Wang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9559-9569
Understanding network generalization and feature discrimination is an open research problem in visual recognition. Many studies have been conducted to assess the quality of feature representations. One of the simple strategies is to utilize a linear probing classifier to quantitatively evaluate the class accuracy under the obtained features. The typical linear probe is only applied as a proxy at the inference time, but its efficacy in measuring features' suitability for linear classification is largely neglected in training. In this paper, we propose an episodic linear probing (ELP) classifier to reflect the generalization of visual representations in an online manner. ELP is trained with detached features from the network and re-initialized episodically. It demonstrates the discriminability of the visual representations in training. Then, an ELP-suitable Regularization term (ELP-SR) is introduced to reflect the distances of probability distributions between ELP classifier and the main classifier. ELP-SR leverages a re-scaling factor to regularize each sample in training, which modulates the loss function adaptively and encourages the features to be discriminative and generalized. We observe significant improvements in three real-world visual recognition tasks, including fine-grained visual classification, long-tailed visual recognition, and generic object recognition. The performance gains show the effectiveness of our method in improving network generalization and feature discrimination.
****
Matching Feature Sets for Few-Shot Image Classification
Arman Afrasiyabi, Hugo Larochelle, Jean-François Lalonde, Christian Gagné; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9014-9024
In image classification, it is common practice to train deep networks to extract a single feature vector per input image. Few-shot classification methods also mostly follow this trend. In this work, we depart from this established direction and instead propose to extract sets of feature vectors for each image. We argue a set-based representation intrinsically builds a richer representation of images from the base classes, which can subsequently better transfer to the few-shot classes. To do so, we propose to adapt existing feature extractors to instead produce sets of feature vectors from images. Our approach, dubbed SetFeat, embeds shallow self-attention mechanisms inside existing encoder architectures. The attention modules are lightweight, and as such our method results in encoders that have approximately the same number of parameters as their original versions. During training and inference, a set-to-set matching metric is used to perform image classification. The effectiveness of our proposed architecture and metrics is demonstrated via thorough experiments on standard few-shot datasets--namely miniImageNet, tieredImageNet, and CUB--in both the 1- and 5-shot scenarios. In all cases but one, our method outperforms the state-of-the-art.
****
DIVeR: Real-Time and Accurate Neural Radiance Fields With Deterministic Integration for Volume Rendering
Liwen Wu, Jae Yong Lee, Anand Bhattad, Yu-Xiong Wang, David Forsyth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16200-16209
DIVeR builds on the key ideas of NeRF and its variants -- density models and volume rendering -- to learn 3D object models that can be rendered realistically from small numbers of images. In contrast to all previous NeRF methods, DIVeR uses deterministic rather than stochastic estimates of the volume rendering integral. DIVeR's representation is a voxel based field of features. To compute the volume rendering integral, a ray is broken into intervals, one per voxel; components of the volume rendering integral are estimated from the features for each interval using an MLP, and the components are aggregated. As a result, DIVeR can render thin translucent structures that are missed by other integrators. Furthermore, DIVeR's representation has semantics that is relatively exposed compared to other such methods -- moving feature vectors around in the voxel space results in

natural edits. Extensive qualitative and quantitative comparisons to current sta
te-of-the-art methods show that DIVeR produces models that (1) render at or abov
e state-of-the-art quality, (2) are very small without being baked, (3) render v
ery fast without being baked, and (4) can be edited in natural ways.
****

Enhancing Classifier Conservativeness and Robustness by Polynomiality
Ziqi Wang, Marco Loog; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2022, pp. 13327-13336
We illustrate the detrimental effect, such as overconfident decisions, that expo
nential behavior can have in methods like classical LDA and logistic regression.
 We then show how polynomiality can remedy the situation. This, among others, le
ads purposefully to random-level performance in the tails, away from the bulk of
 the training data. A directly related, simple, yet important technical novelty
we subsequently present is softRmax: a reasoned alternative to the standard soft
max function employed in contemporary (deep) neural networks. It is derived thro
ugh linking the standard softmax to Gaussian class-conditional models, as employ
ed in LDA, and replacing those by a polynomial alternative. We show that two asp
ects of softRmax, conservativeness and inherent gradient regularization, lead to
 robustness against adversarial attacks without gradient obfuscation.
****

Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic
Segmentation and Localization
Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, Andrea Vedaldi; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
22, pp. 8364-8375
Unsupervised localization and segmentation are long-standing computer vision cha
llenges that involve decomposing an image into semantically-meaningful segments
without any labeled data. These tasks are particularly interesting in an unsuper
vised setting due to the difficulty and cost of obtaining dense image annotation
s, but existing unsupervised approaches struggle with complex scenes containing
multiple objects. Differently from existing methods, which are purely based on d
eep learning, we take inspiration from traditional spectral segmentation methods
 by reframing image decomposition as a graph partitioning problem. Specifically,
 we examine the eigenvectors of the Laplacian of a feature affinity matrix from
self-supervised networks. We find that these eigenvectors already decompose an i
mage into meaningful segments, and can be readily used to localize objects in a
scene. Furthermore, by clustering the features associated with these segments ac
ross a dataset, we can obtain well-delineated, nameable regions, i.e. semantic s
egmentations. Experiments on complex datasets (Pascal VOC, MS-COCO) demonstrate
that our simple spectral method outperforms the state-of-the-art in unsupervised
 localization and segmentation by a significant margin. Furthermore, our method
can be readily used for a variety of complex image editing tasks, such as backgr
ound removal and compositing.
****

OcclusionFusion: Occlusion-Aware Motion Estimation for Real-Time Dynamic 3D Reco
nstruction
Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, Feng Xu; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1736-1745
RGBD-based real-time dynamic 3D reconstruction suffers from inaccurate inter-fra
me motion estimation as errors may accumulate with online tracking. This problem
 is even more severe for single-view-based systems due to strong occlusions. Bas
ed on these observations, we propose OcclusionFusion, a novel method to calculat
e occlusion-aware 3D motion to guide the reconstruction. In our technique, the m
otion of visible regions is first estimated and combined with temporal informati
on to infer the motion of the occluded regions through an LSTM-involved graph ne
ural network. Furthermore, our method computes the confidence of the estimated m
otion by modeling the network output with a probabilistic model, which alleviate
s untrustworthy motions and enables robust tracking. Experimental results on pub
lic datasets and our own recorded data show that our technique outperforms exist
ing single-view-based real-time methods by a large margin. With the reduction of

the motion errors, the proposed technique can handle long and challenging motion sequences. Please check out the project page for sequence results: https://wenbin-lin.github.io/OcclusionFusion.

****

ContIG: Self-Supervised Multimodal Contrastive Learning for Medical Imaging With Genetics

Aiham Taleb, Matthias Kirchler, Remo Monti, Christoph Lippert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20908-20921

High annotation costs are a substantial bottleneck in applying modern deep learning architectures to clinically relevant medical use cases, substantiating the need for novel algorithms to learn from unlabeled data. In this work, we propose ContIG, a self-supervised method that can learn from large datasets of unlabeled medical images and genetic data. Our approach aligns images and several genetic modalities in the feature space using a contrastive loss. We design our method to integrate multiple modalities of each individual person in the same model end-to-end, even when the available modalities vary across individuals. Our procedure outperforms state-of-the-art self-supervised methods on all evaluated downstream benchmark tasks. We also adapt gradient-based explainability algorithms to better understand the learned cross-modal associations between the images and genetic modalities. Finally, we perform genome-wide association studies on the features learned by our models, uncovering interesting relationships between images and genetic data.

****

Revisiting Domain Generalized Stereo Matching Networks From a Feature Consistency Perspective

Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, Edwin R. Hancock; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13001-13011

Despite recent stereo matching networks achieving impressive performance given sufficient training data, they suffer from domain shifts and generalize poorly to unseen domains. We argue that maintaining feature consistency between matching pixels is a vital factor for promoting the generalization capability of stereo matching networks, which has not been adequately considered. Here we address this issue by proposing a simple pixel-wise contrastive learning across the viewpoints. The stereo contrastive feature loss function explicitly constrains the consistency between learned features of matching pixel pairs which are observations of the same 3D points. A stereo selective whitening loss is further introduced to better preserve the stereo feature consistency across domains, which decorrelates stereo features from stereo viewpoint-specific style information. Counter-intuitively, the generalization of feature consistency between two viewpoints in the same scene translates to the generalization of stereo matching performance to unseen domains. Our method is generic in nature as it can be easily embedded into existing stereo networks and does not require access to the samples in the target domain. When trained on synthetic data and generalized to four real-world testing sets, our method achieves superior performance over several state-of-the-art networks.

****

MonoScene: Monocular 3D Semantic Scene Completion

Anh-Quan Cao, Raoul de Charette; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3991-4001

MonoScene proposes a 3D Semantic Scene Completion (SSC) framework, where the dense geometry and semantics of a scene are inferred from a single monocular RGB image. Different from the SSC literature, relying on 2.5 or 3D input, we solve the complex problem of 2D to 3D scene reconstruction while jointly inferring its semantics. Our framework relies on successive 2D and 3D UNets bridged by a novel 2D-3D features projection inspired by optics and introduces a 3D context relation prior to enforce spatio-semantic consistency. Along with architectural contributions, we introduce novel global scene and local frustums losses. Experiments show we outperform the literature on all metrics and datasets while hallucinating

plausible scenery even beyond the camera field of view. Our code and trained models are available at https://github.com/cv-rits/MonoScene.
****

TubeFormer-DeepLab: Video Mask Transformer
Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13914-13924
We present TubeFormer-DeepLab, the first attempt to tackle multiple core video segmentation tasks in a unified manner. Different video segmentation tasks (e.g., video semantic/instance/panoptic segmentation) are usually considered as distinct problems. State-of-the-art models adopted in the separate communities have diverged, and radically different approaches dominate in each task. By contrast, we make a crucial observation that video segmentation tasks could be generally formulated as the problem of assigning different predicted labels to video tubes (where a tube is obtained by linking segmentation masks along the time axis) and the labels may encode different values depending on the target task. The observation motivates us to develop TubeFormer-DeepLab, a simple and effective video mask transformer model that is widely applicable to multiple video segmentation tasks. TubeFormer-DeepLab directly predicts video tubes with task-specific labels (either pure semantic categories, or both semantic categories and instance identities), which not only significantly simplifies video segmentation models, but also advances state-of-the-art results on multiple video segmentation benchmarks.
****

XMP-Font: Self-Supervised Cross-Modality Pre-Training for Few-Shot Font Generation
Wei Liu, Fangyue Liu, Fei Ding, Qian He, Zili Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7905-7914
Generating a new font library is a very labor-intensive and time-consuming job for glyph-rich scripts. Few-shot font generation is thus required, as it requires only a few glyph references without fine-tuning during test. Existing methods follow the style-content disentanglement paradigm, and expect novel fonts to be produced by combining the style codes of the reference glyphs and the content representations of the source. However, these few-shot font generation methods either fail to capture content-independent style representations, or employ localized component-wise style representations, which is insufficient to model many Chinese font styles that involve hyper-component features such as inter-component spacing and "connected-stroke". To resolve these drawbacks and make the style representations more reliable, we propose a self-supervised cross-modality pre-training strategy and a cross-modality transformer-based encoder that is conditioned jointly on the glyph image and the corresponding stroke labels. The cross-modality encoder is pre-trained in a self-supervised manner to allow effective capture of cross- and intra-modality correlations, which facilitates the content-style disentanglement and modeling style representations of all scales (stroke-level, components-level and character-level). The pre-trained encoder is then applied to the downstream font generation task without fine-tuning. Experimental comparisons of our method with state-of-the-art methods demonstrate our method successfully transfers styles of all scales. In addition, it only requires one reference glyph and achieves the lowest rate of bad cases in the few-shot font generation task (28% lower than the second best).
****

Disentangling Visual and Written Concepts in CLIP
Joanna Materzy■ska, Antonio Torralba, David Bau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16410-16419
The CLIP network measures the similarity between natural text and images; in this work, we investigate the entanglement of the representation of word images and natural images in its image encoder. First, we find that the image encoder has an ability to match word images with natural images of scenes described by those words. This is consistent with previous research that suggests that the meaning and the spelling of a word might be entangled deep within the network. On the other hand, we also find that CLIP has a strong ability to match nonsense words,

suggesting that processing of letters is separated from processing of their meaning. To explicitly determine whether the spelling capability of CLIP is separable, we devise a procedure for identifying representation subspaces that selectively isolate or eliminate spelling capabilities. We benchmark our methods against a range of retrieval tasks, and we also test them by measuring the appearance of text in CLIP-guided generated images. We find that our methods are able to cleanly separate spelling capabilities of CLIP from the visual processing of natural images.

****

Gradient-SDF: A Semi-Implicit Surface Representation for 3D Reconstruction

Christiane Sommer, Lu Sang, David Schubert, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6280-6289

We present Gradient-SDF, a novel representation for 3D geometry that combines the advantages of implict and explicit representations. By storing at every voxel both the signed distance field as well as its gradient vector field, we enhance the capability of implicit representations with approaches originally formulated for explicit surfaces. As concrete examples, we show that (1) the Gradient-SDF allows us to perform direct SDF tracking from depth images, using efficient storage schemes like hash maps, and that (2) the Gradient-SDF representation enables us to perform photometric bundle adjustment directly in a voxel representation (without transforming into a point cloud or mesh), naturally a fully implicit optimization of geometry and camera poses and easy geometry upsampling. Experimental results confirm that this leads to significantly sharper reconstructions. Since the overall SDF voxel structure is still respected, the proposed Gradient-SDF is equally suited for (GPU) parallelization as related approaches.

****

Bilateral Video Magnification Filter

Shoichiro Takeda, Kenta Niwa, Mariko Isogawa, Shinya Shimizu, Kazuki Okami, Yushi Aono; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17369-17378

Eulerian video magnification (EVM) has progressed to magnify subtle motions with a target frequency even under the presence of large motions of objects. However, existing EVM methods often fail to produce desirable results in real videos due to (1) mis-extracting subtle motions with a non-target frequency and (2) collapsing results when large de/acceleration motions occur (e.g., objects suddenly start, stop, or change direction). To enhance EVM performance on real videos, this paper proposes a bilateral video magnification filter (BVMF) that offers simple yet robust temporal filtering. BVMF has two kernels; (I) one kernel performs temporal bandpass filtering via a Laplacian of Gaussian whose passband peaks at the target frequency with unity gain and (II) the other kernel excludes large motions outside the magnitude of interest by Gaussian filtering on the intensity of the input signal via the Fourier shift theorem. Thus, BVMF extracts only subtle motions with the target frequency while excluding large motions outside the magnitude of interest, regardless of motion dynamics. In addition, BVMF runs the two kernels in the temporal and intensity domains simultaneously like the bilateral filter does in the spatial and intensity domains. This simplifies implementation and, as a secondary effect, keeps the memory usage low. Experiments conducted on synthetic and real videos show that BVMF outperforms state-of-the-art methods.

****

AdaFocus V2: End-to-End Training of Spatial Dynamic Networks for Video Recognition

Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20062-20072

Recent works have shown that the computational efficiency of video recognition can be significantly improved by reducing the spatial redundancy. As a representative work, the adaptive focus method (AdaFocus) has achieved a favorable trade-off between accuracy and inference speed by dynamically identifying and attending

to the informative regions in each video frame. However, AdaFocus requires a co
mplicated three-stage training pipeline (involving reinforcement learning), lead
ing to slow convergence and is unfriendly to practitioners. This work reformulat
es the training of AdaFocus as a simple one-stage algorithm by introducing a dif
ferentiable interpolation-based patch selection operation, enabling efficient en
d-to-end optimization. We further present an improved training scheme to address
the issues introduced by the one-stage formulation, including the lack of super
vision, input diversity and training stability. Moreover, a conditional-exit tec
hnique is proposed to perform temporal adaptive computation on top of AdaFocus w
ithout additional training. Extensive experiments on six benchmark datasets (i.e
., ActivityNet, FCVID, Mini-Kinetics, Something-Something V1&V2, and Jester) dem
onstrate that our model significantly outperforms the original AdaFocus and othe
r competitive baselines, while being considerably more simple and efficient to t
rain. Code is available at https://github.com/ LeapLabTHU/AdaFocusV2.
****

Localization Distillation for Dense Object Detection
Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, Mi
ng-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2022, pp. 9407-9416
Knowledge distillation (KD) has witnessed its powerful capability in learning co
mpact models in object detection. Previous KD methods for object detection mostl
y focus on imitating deep features within the imitation regions instead of logit
mimicking on classification due to the inefficiency in distilling localization
information. In this paper, by reformulating the knowledge distillation process
on localization, we present a novel localization distillation (LD) method which
can efficiently transfer the localization knowledge from the teacher to the stud
ent. Moreover, we also heuristically introduce the concept of valuable localizat
ion region that can aid to selectively distill the semantic and localization kno
wledge for a certain region. Combining these two new components, for the first t
ime, we show that logit mimicking can outperform feature imitation and, localiza
tion knowledge distillation is more important and efficient than semantic knowle
dge for distilling object detectors. Our distillation scheme is simple as well a
s effective and can be easily applied to different dense object detectors. Exper
iments show that our LD can boost the AP score of GFocal-ResNet-50 with a single
-scale 1x training schedule from 40.1 to 42.1 on the COCO benchmark without any
sacrifice on the inference speed. Our source code and pretrained models will be
made publicly available.
****

What's in Your Hands? 3D Reconstruction of Generic Objects in Hands
Yufei Ye, Abhinav Gupta, Shubham Tulsiani; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3895-3905
Our work aims to reconstruct hand-held objects given a single RGB image. In cont
rast to prior works that typically assume known 3D templates and reduce the prob
lem to 3D pose estimation, our work reconstructs generic hand-held objects witho
ut knowing their 3D templates. Our key insight is that hand articulation is high
ly predictive of the object shape, and we propose an approach that conditionally
reconstructs the object based on the articulation and the visual input. Given a
n image depicting a hand-held object, we first use off-the-shelf systems to esti
mate the underlying hand pose and then infer the object shape in a normalized ha
nd-centric coordinate frame. We parameterized the object by signed distance whic
h is inferred by an implicit network that leverages the information from both vi
sual feature and articulation-aware coordinates to process a query point. We per
form experiments across three datasets and show that our method consistently out
performs baselines and is able to reconstruct a diverse set of objects. We analy
ze the benefits and robustness of explicit articulation conditioning and also sh
ow that this allows the hand pose estimation to further improve in test-time opt
imization.
****

Continuous Scene Representations for Embodied AI
Samir Yitzhak Gadre, Kiana Ehsani, Shuran Song, Roozbeh Mottaghi; Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14849-14859

We propose Continuous Scene Representations (CSR), a scene representation constructed by an embodied agent navigating within a space, where objects and their relationships are modeled by continuous valued embeddings. Our method captures feature relationships between objects, composes them into a graph structure on-the-fly, and situates an embodied agent within the representation. Our key insight is to embed pair-wise relationships between objects in a latent space. This allows for a richer representation compared to discrete relations (e.g., [support], [next-to]) commonly used for building scene representations. CSR can track objects as the agent moves in a scene, update the representation accordingly, and detect changes in room configurations. Using CSR, we outperform state-of-the-art approaches for the challenging downstream task of visual room rearrangement, without any task specific training. Moreover, we show the learned embeddings capture salient spatial details of the scene and show applicability to real world data. A summery video and code is available at https://prior.allenai.org/projects/csr.

****

Beyond 3D Siamese Tracking: A Motion-Centric Paradigm for 3D Single Object Tracking in Point Clouds

Chaoda Zheng, Xu Yan, Haiming Zhang, Baoyuan Wang, Shenghui Cheng, Shuguang Cui, Zhen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8111-8120

3D single object tracking (3D SOT) in LiDAR point clouds plays a crucial role in autonomous driving. Current approaches all follow the Siamese paradigm based on appearance matching. However, LiDAR point clouds are usually textureless and incomplete, which hinders effective appearance matching. Besides, previous methods greatly overlook the critical motion clues among targets. In this work, beyond 3D Siamese tracking, we introduce a motion-centric paradigm to handle 3D SOT from a new perspective. Following this paradigm, we propose a matching-free two-stage tracker M^2-Track. At the 1^st-stage, M^2-Track localizes the target within successive frames via motion transformation. Then it refines the target box through motion-assisted shape completion at the 2^nd-stage. Extensive experiments confirm that M^2-Track significantly outperforms previous state-of-the-arts on three large-scale datasets while running at 57FPS (8%, 17%, and 22%) precision gains on KITTI, NuScenes, and Waymo Open Dataset respectively). Further analysis verifies each component's effectiveness and shows the motion-centric paradigm's promising potential when combined with appearance matching. Code will be made available at https://github.com/Ghostish/Open3DSOT.

****

Neural Mean Discrepancy for Efficient Out-of-Distribution Detection

Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, H.T. Kung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19217-19227

Various approaches have been proposed for out-of-distribution (OOD) detection by augmenting models, input examples, training set, and optimization objectives. Deviating from existing work, we have a simple hypothesis that standard off-the-shelf models may already contain sufficient information about the training set distribution which can be leveraged for reliable OOD detection. Our empirical study on validating this hypothesis, which measures the model activation's mean for OOD and in-distribution (ID) mini-batches, surprisingly finds that activation means of OOD mini-batches consistently deviate more from those of the training data. In addition, training data's activation means can be computed offline efficiently or retrieved from batch normalization layers as a "free lunch". Based upon this observation, we propose a novel metric called Neural Mean Discrepancy (NMD), which compares neural means of the input examples and training data. Leveraging the simplicity of NMD, we propose an efficient OOD detector that computes neural means by a standard forward pass followed by a lightweight classifier. Extensive experiments show that NMD outperforms state-of-the-art OOD approaches across multiple datasets and model architectures in terms of both detection accuracy and computational cost.

****
Non-Probability Sampling Network for Stochastic Human Trajectory Prediction
Inhwan Bae, Jin-Hwi Park, Hae-Gon Jeon; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6477-6487
Capturing multimodal natures is essential for stochastic pedestrian trajectory p
rediction, to infer a finite set of future trajectories. The inferred trajectori
es are based on observation paths and the latent vectors of potential decisions
of pedestrians in the inference step. However, stochastic approaches provide var
ying results for the same data and parameter settings, due to the random samplin
g of the latent vector. In this paper, we analyze the problem by reconstructing
and comparing probabilistic distributions from prediction samples and socially-a
cceptable paths, respectively. Through this analysis, we observe that the infere
nces of all stochastic models are biased toward the random sampling, and fail to
generate a set of realistic paths from finite samples. The problem cannot be re
solved unless an infinite number of samples is available, which is infeasible in
practice. We introduce that the Quasi-Monte Carlo (QMC) method, ensuring unifor
m coverage on the sampling space, as an alternative to the conventional random s
ampling. With the same finite number of samples, the QMC improves all the multim
odal prediction results. We take an additional step ahead by incorporating a lea
rnable sampling network into the existing networks for trajectory prediction. Fo
r this purpose, we propose the Non-Probability Sampling Network (NPSN), a very s
mall network (5K parameters) that generates purposive sample sequences using th
e past paths of pedestrians and their social interactions. Extensive experiments
confirm that NPSN can significantly improve both the prediction accuracy (up to
60%) and reliability of the public pedestrian trajectory prediction benchmark.
Code is publicly available at https://github.com/inhwanbae/NPSN.
****
Marginal Contrastive Correspondence for Guided Image Generation
Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Changgong Zh
ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2022, pp. 10663-10672
Exemplar-based image translation establishes dense correspondences between a con
ditional input and an exemplar (from two different domains) for leveraging detai
led exemplar styles to achieve realistic image translation. Existing work builds
the cross-domain correspondences implicitly by minimizing feature-wise distance
s across the two domains. Without explicit exploitation of domain-invariant feat
ures, this approach may not reduce the domain gap effectively which often leads
to sub-optimal correspondences and image translation. We design a Marginal Contr
astive Learning Network (MCL-Net) that explores contrastive learning to learn do
main-invariant features for realistic exemplar-based image translation. Specific
ally, we design an innovative marginal contrastive loss that guides to establish
dense correspondences explicitly. Nevertheless, building correspondence with do
main-invariant semantics alone may impair the texture patterns and lead to degra
ded texture generation. We thus design a Self-Correlation Map (SCM) that incorpo
rates scene structures as auxiliary information which improves the built corresp
ondences substantially. Quantitative and qualitative experiments on multifarious
image translation tasks show that the proposed method outperforms the state-of-
the-art consistently.
****
Complex Backdoor Detection by Symmetric Feature Differencing
Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, Xiangyu Zhang
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2022, pp. 15003-15013
Many existing backdoor scanners work by finding a small and fixed trigger. Howev
er, advanced attacks have large and pervasive triggers, rendering existing scann
ers less effective. We develop a new detection method. It first uses a trigger i
nversion technique to generate triggers, namely, universal input patterns flippi
ng victim class samples to a target class. It then checks if any such trigger is
composed of features that are not natural distinctive features between the vict
im and target classes. It is based on a novel symmetric feature differencing met

hod that identifies features separating two sets of samples (e.g., from two respective classes). We evaluate the technique on a number of advanced attacks including composite attack, reflection attack, hidden attack, filter attack, and also on the traditional patch attack. The evaluation is on thousands of models, including both clean and trojaned models, with various architectures. We compare with three state-of-the-art scanners. Our technique can achieve 80-88% accuracy while the baselines can only achieve 50-70% on complex attacks. Our results on the TrojAI competition rounds 2-4, which have patch backdoors and filter backdoors, show that existing scanners may produce hundreds of false positives (i.e., clean models recognized as trojaned), while our technique removes 78-100% of them with a small increase of false negatives by 0-30%, leading to 17-41% overall accuracy improvement. This allows us to achieve top performance on the leaderboard.

****

Time Lens++: Event-Based Frame Interpolation With Parametric Non-Linear Flow and Multi-Scale Fusion

Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17755-17764

Recently, video frame interpolation using a combination of frame- and event-based cameras has surpassed traditional image-based methods both in terms of performance and memory efficiency. However, current methods still suffer from (i) brittle image-level fusion of complementary interpolation results, that fails in the presence of artifacts in the fused image, (ii) potentially temporally inconsistent and inefficient motion estimation procedures, that run for every inserted frame and (iii) low contrast regions that do not trigger events, and thus cause events-only motion estimation to generate artifacts. Moreover, previous methods were only tested on datasets consisting of planar and far-away scenes, which do not capture the full complexity of the real world. In this work, we address the above problems by introducing multi-scale feature-level fusion and computing one-shot non-linear inter-frame motion---which can be efficiently sampled for image warping---from events and images. We also collect the first large-scale events and frames dataset consisting of more than 100 challenging scenes with depth variations, captured with a new experimental setup based on a beamsplitter. We show that our method improves the reconstruction quality by up to 0.2 dB in terms of PSNR and by up to 15% in LPIPS score. Code and dataset will be released upon acceptance.

****

ResSFL: A Resistance Transfer Framework for Defending Model Inversion Attack in Split Federated Learning

Jingtao Li, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, Chaitali Chakrabarti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10194-10202

This work aims to tackle Model Inversion (MI) attack on Split Federated Learning (SFL). SFL is a recent distributed training scheme where multiple clients send intermediate activations (i.e., feature map), instead of raw data, to a central server. While such a scheme helps reduce the computational load at the client end, it opens itself to reconstruction of raw data from intermediate activation by the server. Existing works on protecting SFL only consider inference and do not handle attacks during training. So we propose ResSFL, a Split Federated Learning Framework that is designed to be MI-resistant during training. It is based on deriving a resistant feature extractor via attacker-aware training, and using this extractor to initialize the client-side model prior to standard SFL training. Such a method helps in reducing the computational complexity due to use of strong inversion model in client-side adversarial training as well as vulnerability of attacks launched in early training epochs. On CIFAR-100 dataset, our proposed framework successfully mitigates MI attack on a VGG-11 model with a high reconstruction Mean-Square-Error of 0.050 compared to 0.005 obtained by the baseline system. The framework achieves 67.5% accuracy (only 1% accuracy drop) with very low computation overhead. Code is released at: https://github.com/zlijingtao/ResSFL

****
RecDis-SNN: Rectifying Membrane Potential Distribution for Directly Training Spiking Neural Networks
Yufei Guo, Xinyi Tong, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Zhe Ma, Xuhui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 326-335
The brain-inspired and event-driven Spiking Neural Network (SNN) aims at mimicking the synaptic activity of biological neurons, which transmits binary spike signals between network units when the membrane potential exceeds the firing threshold. This bio-mimetic mechanism of SNN appears energy-efficiency with its power sparsity and asynchronous operations on spike events. Unfortunately, with the propagation of binary spikes, the distribution of membrane potential will shift, leading to degeneration, saturation, and gradient mismatch problems, which would be disadvantageous to the network optimization and convergence. Such undesired shifts would prevent the SNN from performing well and going deep. To tackle these problems, we attempt to rectify the membrane potential distribution (MPD) by designing a novel distribution loss, MPD-Loss, which can explicitly penalize the undesired shifts without introducing any additional operations in the inference phase. Moreover, the proposed method can also mitigate the quantization error in SNNs, which is usually ignored in other works. Experimental results demonstrate that the proposed method can directly train a deeper, larger and better performing SNN within fewer timesteps.
****
Human-Aware Object Placement for Visual Environment Reconstruction
Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3959-3970
Humans are in constant contact with the world as they move through it and interact with it. This contact is a vital source of information for understanding 3D humans, 3D scenes, and the interactions between them. In fact, we demonstrate that these human-scene interactions (HSIs) can be leveraged to improve the 3D reconstruction of a scene from a monocular RGB video. Our key idea is that, as a person moves through a scene and interacts with it, we accumulate HSIs across multiple input images, and use these in optimizing the 3D scene to reconstruct a consistent, physically plausible, 3D scene layout. Our optimization-based approach exploits three types of HSI constraints: (1) humans who move in a scene are occluded by, or occlude, objects, thus constraining the depth ordering of the objects, (2) humans move through free space and do not interpenetrate objects, (3) when humans and objects are in contact, the contact surfaces occupy the same place in space. Using these constraints in an optimization formulation across all observations, we significantly improve 3D scene layout reconstruction. Furthermore, we show that our scene reconstruction can be used to refine the initial 3D human pose and shape (HPS) estimation. We evaluate the 3D scene layout reconstruction and HPS estimates qualitatively and quantitatively using the PROX and PiGraphs datasets. The code and data are available for research purposes at https://mover.is.tue.mpg.de.
****
X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval
Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, Guangwei Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5006-5015
In text-video retrieval, the objective is to learn a cross-modal similarity function between a text and a video that ranks relevant text-video pairs higher than irrelevant pairs. However, videos inherently express a much wider gamut of information than texts. Instead, texts often capture sub-regions of entire videos and are most semantically similar to certain frames within videos. Therefore, for a given text, a retrieval model should focus on the text's most semantically similar video sub-regions to make a more relevant comparison. Yet, most existing works aggregate entire videos without directly considering text. Common text-agnostic aggregations schemes include mean-pooling or self-attention over the frames,

but these are likely to encode misleading visual information not described in the given text. To address this, we propose a cross-modal attention model called X-Pool that reasons between a text and the frames of a video. Our core mechanism is a scaled dot product attention for a text to attend to its most semantically similar frames. We then generate an aggregated video representation conditioned on the text's attention weights over the frames. We evaluate our method on three benchmark datasets of MSR-VTT, MSVD and LSMDC, achieving new state-of-the-art results by up to 12% in relative improvement in Recall@1. Our findings thereby highlight the importance of joint text-video reasoning to extract important visual cues according to text. Full code and demo can be found at: https://layer6ai-labs.github.io/xpool/

****

Learning of Global Objective for Network Flow in Multi-Object Tracking
Shuai Li, Yu Kong, Hamid Rezatofighi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8855-8865
This paper concerns the problem of multi-object tracking based on the min-cost flow (MCF) formulation, which is conventionally studied as an instance of linear program. Given its computationally tractable inference, the success of MCF tracking largely relies on the learned cost function of underlying linear program. Most previous studies focus on learning the cost function by only taking into account two frames during training, therefore the learned cost function is sub-optimal for MCF where a multi-frame data association must be considered during inference. In order to address this problem, in this paper we propose a novel differentiable framework that ties training and inference together during learning by solving a bi-level optimization problem, where the lower-level solves a linear program and the upper-level contains a loss function that incorporates global tracking result. By back-propagating the loss through differentiable layers via gradient descent, the globally parameterized cost function is explicitly learned and regularized. With this approach, we are able to learn a better objective for global MCF tracking. As a result, we achieve competitive performances compared to the current state-of-the-art methods on the popular multi-object tracking benchmarks such as MOT16, MOT17 and MOT20

****

Towards Weakly-Supervised Text Spotting Using a Multi-Task Transformer
Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R. Manmatha, Pietro Perona; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4604-4613
Text spotting end-to-end methods have recently gained attention in the literature due to the benefits of jointly optimizing the text detection and recognition components. Existing methods usually have a distinct separation between the detection and recognition branches, requiring exact annotations for the two tasks. We introduce TextTranSpotter (TTS), a transformer-based approach for text spotting and the first text spotting framework which may be trained with both fully- and weakly-supervised settings. By learning a single latent representation per word detection, and using a novel loss function based on the Hungarian loss, our method alleviates the need for expensive localization annotations. Trained with only text transcription annotations on real data, our weakly-supervised method achieves competitive performance with previous state-of-the-art fully-supervised methods. When trained in a fully-supervised manner, TextTranSpotter shows state-of-the-art results on multiple benchmarks.

****

Gated2Gated: Self-Supervised Depth Estimation From Gated Images
Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2811-2821
Gated cameras hold promise as an alternative to scanning LiDAR sensors with high-resolution 3D depth that is robust to back-scatter in fog, snow, and rain. Instead of sequentially scanning a scene and directly recording depth via the photon time-of-flight, as in pulsed LiDAR sensors, gated imagers encode depth in the relative intensity of a handful of gated slices, captured at megapixel resolution

. Although existing methods have shown that it is possible to decode high-resolu
tion depth from such measurements, these methods require synchronized and calibr
ated LiDAR to supervise the gated depth decoder - prohibiting fast adoption acro
ss geographies, training on large unpaired datasets, and exploring alternative a
pplications outside of automotive use cases. In this work, we fill this gap and
propose an entirely self-supervised depth estimation method that uses gated inte
nsity profiles and temporal consistency as a training signal. The proposed model
 is trained end-to-end from gated video sequences, does not require LiDAR or RGB
 data, and learns to estimate absolute depth values. We take gated slices as inp
ut and disentangle the estimation of the scene albedo, depth, and ambient light,
 which are then used to learn to reconstruct the input slices through a cyclic l
oss. We rely on temporal consistency between a given frame and neighboring gated
 slices to estimate depth in regions with shadows and reflections. We experiment
ally validate that the proposed approach outperforms existing supervised and sel
f-supervised depth estimation methods based on monocular RGB and stereo images,
as well as supervised methods based on gated images.
****

RAMA: A Rapid Multicut Algorithm on GPU
Ahmed Abbas, Paul Swoboda; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2022, pp. 8193-8202
We propose a highly parallel primal-dual algorithm for the multicut (a.k.a. corr
elation clustering) problem, a classical graph clustering problem widely used in
 machine learning and computer vision. Our algorithm consists of three steps exe
cuted recursively: (1) Finding conflicted cycles that correspond to violated ine
qualities of the underlying multicut relaxation, (2) Performing message passing
between the edges and cycles to optimize the Lagrange relaxation coming from the
 found violated cycles producing reduced costs and (3) Contracting edges with hi
gh reduced costs through matrix-matrix multiplications. Our algorithm produces p
rimal solutions and lower bounds that estimate the distance to optimum. We imple
ment our algorithm on GPUs and show resulting one to two orders-of-magnitudes im
provements in execution speed without sacrificing solution quality compared to t
raditional sequential algorithms that run on CPUs. We can solve very large scale
 benchmark problems with up to O(10^8) variables in a few seconds with small pri
mal-dual gaps. Our code is available at https://github.com/pawelswoboda/RAMA.
****

Adversarial Parametric Pose Prior
Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salz
mann, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2022, pp. 10997-11005
The Skinned Multi-Person Linear (SMPL) model represents human bodies by mapping
pose and shape parameters to body meshes. However, not all pose and shape parame
ter values yield physically-plausible or even realistic body meshes. In other wo
rds, SMPL is under-constrained and may yield invalid results. We propose learnin
g a prior that restricts the SMPL parameters to values that produce realistic po
ses via adversarial training. We show that our learned prior covers the diversit
y of the real-data distribution, facilitates optimization for 3D reconstruction
from 2D keypoints, and yields better pose estimates when used for regression fro
m images. For all these tasks, it outperforms the state-of-the-art VAE-based app
roach to constraining the SMPL parameters. The code will be made available at ht
tps://github.com/cvlab-epfl/adv_param_pose_prior.
****

DC-SSL: Addressing Mismatched Class Distribution in Semi-Supervised Learning
Zhen Zhao, Luping Zhou, Yue Duan, Lei Wang, Lei Qi, Yinghuan Shi; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
, pp. 9757-9765
Consistency-based Semi-supervised learning (SSL) has achieved promising performa
nce recently. However, the success largely depends on the assumption that the la
beled and unlabeled data share an identical class distribution, which is hard to
 meet in real practice. The distribution mismatch between the labeled and unlabe
led sets can cause severe bias in the pseudo-labels of SSL, resulting in signifi

cant performance degradation. To bridge this gap, we put forward a new SSL learn ing framework, named Distribution Consistency SSL (DC-SSL), which rectifies the pseudo-labels from a distribution perspective. The basic idea is to directly est imate a reference class distribution (RCD), which is regarded as a surrogate of the ground truth class distribution about the unlabeled data, and then improve t he pseudo-labels by encouraging the predicted class distribution (PCD) of the un labeled data to approach RCD gradually. To this end, this paper revisits the Exp onentially Moving Average (EMA) model and utilizes it to estimate RCD in an iter atively improved manner, which is achieved with a momentum-update scheme through out the training procedure. On top of this, two strategies are proposed for RCD to rectify the pseudo-label prediction, respectively. They correspond to an effi cient training-free scheme and a training-based alternative that generates more accurate and reliable predictions. DC-SSL is evaluated on multiple SSL benchmark s and demonstrates remarkable performance improvement over competitive methods u nder matched- and mismatched-distribution scenarios.

****

Mask Transfiner for High-Quality Instance Segmentation

Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2022, pp. 4412-4421

Two-stage and query-based instance segmentation methods have achieved remarkable results. However, their segmented masks are still very coarse. In this paper, w e present Mask Transfiner for high-quality and efficient instance segmentation. Instead of operating on regular dense tensors, our Mask Transfiner decomposes an d represents the image regions as a quadtree. Our transformer-based approach onl y processes detected error-prone tree nodes and self-corrects their errors in pa rallel. While these sparse pixels only constitute a small proportion of the tota l number, they are critical to the final mask quality. This allows Mask Transfin er to predict highly accurate instance masks, at a low computational cost. Exten sive experiments demonstrate that Mask Transfiner outperforms current instance s egmentation methods on three popular benchmarks, significantly improving both tw o-stage and query-based frameworks by a large margin of +3.0 mask AP on COCO and BDD100K, and +6.6 boundary AP on Cityscapes. Our code and trained models are av ailable at https://github.com/SysCV/transfiner.

****

End-to-End Reconstruction-Classification Learning for Face Forgery Detection

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, Xiaokang Yang; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2022, pp. 4113-4122

Existing face forgery detectors mainly focus on specific forgery patterns like n oise characteristics, local textures, or frequency statistics for forgery detect ion. This causes specialization of learned representations to known forgery patt erns presented in the training set, and makes it difficult to detect forgeries w ith unknown patterns. In this paper, from a new perspective, we propose a forger y detection framework emphasizing the common compact representations of genuine faces based on reconstruction-classification learning. Reconstruction learning o ver real images enhances the learned representations to be aware of forgery patt erns that are even unknown, while classification learning takes the charge of mi ning the essential discrepancy between real and fake images, facilitating the un derstanding of forgeries. To achieve better representations, instead of only usi ng the encoder in reconstruction learning, we build bipartite graphs over the en coder and decoder features in a multi-scale fashion. We further exploit the reco nstruction difference as guidance of forgery traces on the graph output as the f inal representation, which is fed into the classifier for forgery detection. The reconstruction and classification learning is optimized end-to-end. Extensive e xperiments on large-scale benchmark datasets demonstrate the superiority of the proposed method over state of the arts.

****

It Is Okay To Not Be Okay: Overcoming Emotional Bias in Affective Image Captioni ng by Contrastive Data Collection

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21263-21272

Datasets that capture the connection between vision, language, and affection are limited, causing a lack of understanding of the emotional aspect of human intelligence. As a step in this direction, the ArtEmis dataset was recently introduced as a large-scale dataset of emotional reactions to images along with language explanations of these chosen emotions. We observed a significant emotional bias towards instance-rich emotions, making trained neural speakers less accurate in describing under-represented emotions. We show that collecting new data, in the same way, is not effective in mitigating this emotional bias. To remedy this problem, we propose a contrastive data collection approach to balance ArtEmis with a new complementary dataset such that a pair of similar images have contrasting emotions (one positive and one negative). We collected 260,533 instances using the proposed method, we combine them with ArtEmis, creating a second iteration of the dataset. The new combined dataset, dubbed ArtEmis v2.0, has a balanced distribution of emotions with explanations revealing more fine details in the associated painting. Our experiments show that neural speakers trained on the new dataset improve CIDEr and METEOR evaluation metrics by 20% and 7%, respectively, compared to the biased dataset. Finally, we also show that the performance per emotion of neural speakers is improved across all the emotion categories, significantly on under-represented emotions. The collected dataset and code are available at https://artemisdataset-v2.org.

****

Transferability Metrics for Selecting Source Model Ensembles

Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7936-7946

We address the problem of ensemble selection in transfer learning: Given a large pool of source models we want to select an ensemble of models which, after fine-tuning on the target training set, yields the best performance on the target test set. Since fine-tuning all possible ensembles is computationally prohibitive, we aim at predicting performance on the target dataset using a computationally efficient transferability metric. We propose several new transferability metrics designed for this task and evaluate them in a challenging and realistic transfer learning setup for semantic segmentation: we create a large and diverse pool of source models by considering 17 source datasets covering a wide variety of image domain, two different architectures, and two pre-training schemes. Given this pool, we then automatically select a subset to form an ensemble performing well on a given target dataset. We compare the ensemble selected by our method to two baselines which select a single source model, either (1) from the same pool as our method; or (2) from a pool containing large source models, each with similar capacity as an ensemble. Averaged over 17 target datasets, we outperform these baselines by 6.0% and 2.5% relative mean IoU, respectively.

****

Neural Global Shutter: Learn To Restore Video From a Rolling Shutter Camera With Global Reset Feature

Zhixiang Wang, Xiang Ji, Jia-Bin Huang, Shin'ichi Satoh, Xiao Zhou, Yinqiang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17794-17803

Most computer vision systems assume distortion-free images as inputs. The widely used rolling-shutter (RS) image sensors, however, suffer from geometric distortion when the camera and object undergo motion during capture. Extensive researches have been conducted on correcting RS distortions. However, most of the existing work relies heavily on the prior assumptions of scenes or motions. Besides, the motion estimation steps are either oversimplified or computationally inefficient due to the heavy flow warping, limiting their applicability. In this paper, we investigate using rolling shutter with a global reset feature (RSGR) to restore clean global shutter (GS) videos. This feature enables us to turn the rectification problem into a deblur-like one, getting rid of inaccurate and costly expl

icit motion estimation. First, we build an optic system that captures paired RSG R/GS videos. Second, we develop a novel algorithm incorporating spatial and temp oral designs to correct the spatial-varying RSGR distortion. Third, we demonstra te that existing image-to-image translation algorithms can recover clean GS vide os from distorted RSGR inputs, yet our algorithm achieves the best performance w ith the specific designs. Our rendered results are not only visually appealing b ut also beneficial to downstream tasks. Compared to the state-of-the-art RS solu tion, our RSGR solution is superior in both effectiveness and efficiency. Consid ering it is easy to realize without changing the hardware, we believe our RSGR s olution can potentially replace the RS solution in taking distortion-free videos with low noise and low budget.

****

DiRA: Discriminative, Restorative, and Adversarial Learning for Self-Supervised Medical Image Analysis

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, Jianming Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 20824-20834

Discriminative learning, restorative learning, and adversarial learning have pro ven beneficial for self-supervised learning schemes in computer vision and medic al imaging. Existing efforts, however, omit their synergistic effects on each ot her in a ternary setup, which, we envision, can significantly benefit deep seman tic representation learning. To realize this vision, we have developed DiRA, the first framework that unites discriminative, restorative, and adversarial learni ng in a unified manner to collaboratively glean complementary visual information from unlabeled medical images for fine-grained semantic representation learning . Our extensive experiments demonstrate that DiRA (1) encourages collaborative l earning among three learning ingredients, resulting in more generalizable repres entation across organs, diseases, and modalities; (2) outperforms fully supervis ed ImageNet models and increases robustness in small data regimes, reducing anno tation cost across multiple medical imaging applications; (3) learns fine-graine d semantic representation, facilitating accurate lesion localization with only i mage-level annotation; and (4) enhances state-of-the-art restorative approaches, revealing that DiRA is a general mechanism for united representation learning. All code and pretrained models are available at https://github.com/JLiangLab/DiR A.

****

Open Challenges in Deep Stereo: The Booster Dataset

Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattocc ia, Luigi Di Stefano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21168-21178

We present a novel high-resolution and challenging stereo dataset framing indoor scenes annotated with dense and accurate ground-truth disparities. Peculiar to our dataset is the presence of several specular and transparent surfaces, i.e. t he main causes of failures for state-of-the-art stereo networks. Our acquisition pipeline leverages a novel deep space-time stereo framework which allows for ea sy and accurate labeling with sub-pixel precision. We release a total of 419 sam ples collected in 64 different scenes and annotated with dense ground-truth disp arities. Each sample include a high-resolution pair (12 Mpx) as well as an unbal anced pair (Left: 12 Mpx, Right: 1.1 Mpx). Additionally, we provide manually ann otated material segmentation masks and 15K unlabeled samples. We evaluate state- of-the-art deep networks based on our dataset, highlighting their limitations in addressing the open challenges in stereo and drawing hints for future research.

****

Location-Free Human Pose Estimation

Xixia Xu, Yingguo Gao, Ke Yan, Xue Lin, Qi Zou; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13137-13146

Human pose estimation (HPE) usually requires large-scale training data to reach high performance. However, it is rather time-consuming to collect high-quality a nd fine-grained annotations for human body. To alleviate this issue, we revisit HPE and propose a location-free framework without supervision of keypoint locati

ons. We reformulate the regression-based HPE from the perspective of classification. Inspired by the CAM-based weakly-supervised object localization, we observe that the coarse keypoint locations can be acquired through the part-aware CAMs but unsatisfactory due to the gap between the fine-grained HPE and the object-level localization. To this end, we propose a customized transformer framework to mine the fine-grained representation of human context, equipped with the structural relation to capture subtle differences among keypoints. Concretely, we design a Multi-scale Spatial-guided Context Encoder to fully capture the global human context while focusing on the part-aware regions and a Relation-encoded Pose Prototype Generation module to encode the structural relations. All these works together for strengthening the weak supervision from image-level category labels on locations. Our model achieves competitive performance on three datasets when only supervised at a category-level and importantly, it can achieve comparable results with fully-supervised methods with only 25% location labels on MS-COCO and MPII.

****

Self-Supervised Bulk Motion Artifact Removal in Optical Coherence Tomography Angiography

Jiaxiang Ren, Kicheon Park, Yingtian Pan, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20617-20625

Optical coherence tomography angiography (OCTA) is an important imaging modality in many bioengineering tasks. The image quality of OCTA, however, is often degraded by Bulk Motion Artifacts (BMA), which are due to micromotion of subjects and typically appear as bright stripes surrounded by blurred areas. State-of-the-art methods usually treat BMA removal as a learning-based image inpainting problem, but require numerous training samples with nontrivial annotation. In addition, these methods discard the rich structural and appearance information carried in the BMA stripe region. To address these issues, in this paper we propose a self-supervised content-aware BMA removal model. First, the gradient-based structural information and appearance feature are extracted from the BMA area and injected into the model to capture more connectivity. Second, with easily collected defective masks, the model is trained in a self-supervised manner, in which only the clear areas are used for training while the BMA areas for inference. With the structural information and appearance feature from noisy image as references, our model can remove larger BMA and produce better visualizing result. In addition, only 2D images with defective masks are involved, hence improving the efficiency of our method. Experiments on OCTA of mouse cortex demonstrate that our model can remove most BMA with extremely large sizes and inconsistent intensities while previous methods fail.

****

Watch It Move: Unsupervised Discovery of 3D Joints for Re-Posing of Articulated Objects

Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, Orazio Gallo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3677-3687

Rendering articulated objects while controlling their poses is critical to applications such as virtual reality or animation for movies. Manipulating the pose of an object, however, requires the understanding of its underlying structure, that is, its joints and how they interact with each other. Unfortunately, assuming the structure to be known, as existing methods do, precludes the ability to work on new object categories. We propose to learn both the appearance and the structure of previously unseen articulated objects by observing them move from multiple views, with no joints annotation supervision, or information about the structure. We observe that 3D points that are static relative to one another should belong to the same part, and that adjacent parts that move relative to each other must be connected by a joint. To leverage this insight, we model the object parts in 3D as ellipsoids, which allows us to identify joints. We combine this explicit representation with an implicit one that compensates for the approximation introduced. We show that our method works for different structures, from quadrup

eds, to single-arm robots, to humans.
****
PoseTrack21: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking
Andreas Döring, Di Chen, Shanshan Zhang, Bernt Schiele, Jürgen Gall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20963-20972
Current research evaluates person search, multi-object tracking and multi-person pose estimation as separate tasks and on different datasets although these tasks are very akin to each other and comprise similar sub-tasks, e.g. person detection or appearance-based association of detected persons. Consequently, approaches on these respective tasks are eligible to complement each other. Therefore, we introduce PoseTrack21, a large-scale dataset for person search, multi-object tracking and multi-person pose tracking in real-world scenarios with a high diversity of poses. The dataset provides rich annotations like human pose annotations including annotations of joint occlusions, bounding box annotations even for small persons, and person-ids within and across video sequences. The dataset allows to evaluate multi-object tracking and multi-person pose tracking jointly with person re-identification or exploit structural knowledge of human poses to improve person search and tracking, particularly in the context of severe occlusions. With PoseTrack21, we want to encourage researchers to work on joint approaches that perform reasonably well on all three tasks.
****
Event-Based Video Reconstruction via Potential-Assisted Spiking Neural Network
Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3594-3604
Neuromorphic vision sensor is a new bio-inspired imaging paradigm that reports asynchronous, continuously per-pixel brightness changes called 'events' with high temporal resolution and high dynamic range. So far, the event-based image reconstruction methods are based on artificial neural networks (ANN) or hand-crafted spatiotemporal smoothing techniques. In this paper, we first implement the image reconstruction work via deep spiking neural network (SNN) architecture. As the bio-inspired neural networks, SNNs operating with asynchronous binary spikes distributed over time, can potentially lead to greater computational efficiency on event-driven hardware. We propose a novel Event-based Video reconstruction framework based on a fully Spiking Neural Network (EVSNN), which utilizes Leaky-Integrate-and-Fire (LIF) neuron and Membrane Potential (MP) neuron. We find that the spiking neurons have the potential to store useful temporal information (memory) to complete such time-dependent tasks. Furthermore, to better utilize the temporal information, we propose a hybrid potential-assisted framework (PA-EVSNN) using the membrane potential of spiking neuron. The proposed neuron is referred as Adaptive Membrane Potential (AMP) neuron, which adaptively updates the membrane potential according to the input spikes. The experimental results demonstrate that our models achieve comparable performance to ANN-based models on IJRR, MVSEC, and HQF datasets. The energy consumptions of EVSNN and PA-EVSNN are 19.36 times and 7.75 times more computationally efficient than their ANN architectures, respectively. The code and pretrained model are available at https://sites.google.com/view/evsnn.
****
Efficient Maximal Coding Rate Reduction by Variational Forms
Christina Baek, Ziyang Wu, Kwan Ho Ryan Chan, Tianjiao Ding, Yi Ma, Benjamin D. Haeffele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 500-508
The principle of Maximal Coding Rate Reduction (MCR2) has recently been proposed as a training objective for learning discriminative low-dimensional structures intrinsic to high-dimensional data to allow for more robust training than standard approaches, such as cross-entropy minimization. However, despite the advantages that have been shown for MCR2 training, MCR2 suffers from a significant computational cost due to the need to evaluate and differentiate a significant number

of log-determinant terms that grows linearly with the number of classes. By tak ing advantage of variational forms of spectral functions of a matrix, we reformu late the MCR2 objective to a form that can scale significantly without compromis ing training accuracy. Experiments in image classification demonstrate that our proposed formulation results in a significant speed up over optimizing the origi nal MCR2 objective directly and often results in higher quality learned represen tations. Further, our approach may be of independent interest in other models th at require computation of log-determinant forms, such as in system identificatio n or normalizing flow models.
****

Ithaca365: Dataset and Driving Perception Under Repeated and Challenging Weather Conditions

Carlos A. Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Mon ica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, Wei-Lun Chao, Bharath Hariha ran, Kilian Q. Weinberger, Mark Campbell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21383-21392

Advances in perception for self-driving cars have accelerated in recent years du e to the availability of large-scale datasets, typically collected at specific l ocations and under nice weather conditions. Yet, to achieve the high safety requ irement, these perceptual systems must operate robustly under a wide variety of weather conditions including snow and rain. In this paper, we present a new data set to enable robust autonomous driving via a novel data collection process --- data is repeatedly recorded along a 15 km route under diverse scene (urban, high way, rural, campus), weather (snow, rain, sun), time (day/night), and traffic co nditions (pedestrians, cyclists and cars). The dataset includes images and point clouds from cameras and LiDAR sensors, along with high-precision GPS/INS to est ablish correspondence across routes. The dataset includes road and object annota tions using amodal masks to capture partial occlusions and 2D/3D bounding boxes. We demonstrate the uniqueness of this dataset by analyzing the performance of b aselines in amodal segmentation of road and objects, depth estimation, and 3D ob ject detection. The repeated routes opens new research directions in object disc overy, continual learning, and anomaly detection.
****

AutoLoss-GMS: Searching Generalized Margin-Based Softmax Loss Function for Perso n Re-Identification

Hongyang Gu, Jianmin Li, Guangyuan Fu, Chifong Wong, Xinghao Chen, Jun Zhu; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2022, pp. 4744-4753

Person re-identification is a hot topic in computer vision, and the loss functio n plays a vital role in improving the discrimination of the learned features. Ho wever, most existing models utilize the hand-crafted loss functions, which are u sually sub-optimal and challenging to be designed. In this paper, we propose a n ovel method, AutoLoss-GMS, to search the better loss function in the space of ge neralized margin-based softmax loss function for person re-identification automa tically. Specifically, the generalized margin-based softmax loss function is fir st decomposed into two computational graphs and a constant. Then a general searc hing framework built upon the evolutionary algorithm is proposed to search for t he loss function efficiently. The computational graph is constructed with a forw ard method, which can construct much richer loss function forms than the backwar d method used in existing works. In addition to the basic in-graph mutation oper ations, the cross-graph mutation operation is designed to further improve the of fspring's diversity. The loss-rejection protocol, equivalence-check strategy and the predictor-based promising-loss chooser are developed to improve the search efficiency. Finally, experimental results demonstrate that the searched loss fun ctions can achieve state-of-the-art performance and be transferable across diffe rent models and datasets in person re-identification.
****

YouMVOS: An Actor-Centric Multi-Shot Video Object Segmentation Dataset

Donglai Wei, Siddhant Kharbanda, Sarthak Arora, Roshan Roy, Nishant Jain, Akash Palrecha, Tanav Shah, Shray Mathur, Ritik Mathur, Abhijay Kemkar, Anirudh Chakra

varthy, Zudi Lin, Won-Dong Jang, Yansong Tang, Song Bai, James Tompkin, Philip H.S. Torr, Hanspeter Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21044-21053
Many video understanding tasks require analyzing multi-shot videos, but existing datasets for video object segmentation (VOS) only consider single-shot videos. To address this challenge, we collected a new dataset---YouMVOS---of 200 popular YouTube videos spanning ten genres, where each video is on average five minutes long and with 75 shots. We selected recurring actors and annotated 431K segmentation masks at a frame rate of six, exceeding previous datasets in average video duration, object variation, and narrative structure complexity. We incorporated good practices of model architecture design, memory management, and multi-shot tracking into an existing video segmentation method to build competitive baseline methods. Through error analysis, we found that these baselines still fail to cope with cross-shot appearance variation on our YouMVOS dataset. Thus, our dataset poses new challenges in multi-shot segmentation towards better video analysis. Data, code, and pre-trained models are available at https://donglaiw.github.io/proj/youMVOS
****

DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation
Lukas Hoyer, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9924-9935
As acquiring pixel-wise annotations of real-world images for semantic segmentation is a costly process, a model can instead be trained with more accessible synthetic data and adapted to real images without requiring their annotations. This process is studied in unsupervised domain adaptation (UDA). Even though a large number of methods propose new adaptation strategies, they are mostly based on outdated network architectures. As the influence of recent network architectures has not been systematically studied, we first benchmark different network architectures for UDA and newly reveal the potential of Transformers for UDA semantic segmentation. Based on the findings, we propose a novel UDA method, DAFormer. The network architecture of DAFormer consists of a Transformer encoder and a multi-level context-aware feature fusion decoder. It is enabled by three simple but crucial training strategies to stabilize the training and to avoid overfitting to the source domain: While (1) Rare Class Sampling on the source domain improves the quality of the pseudo-labels by mitigating the confirmation bias of self-training toward common classes, (2) a Thing-Class ImageNet Feature Distance and (3) a learning rate warmup promote feature transfer from ImageNet pretraining. DAFormer represents a major advance in UDA. It improves the state of the art by 10.8 mIoU for GTA-to-Cityscapes and 5.4 mIoU for Synthia-to-Cityscapes and enables learning even difficult classes such as train, bus, and truck well. The implementation is available at https://github.com/lhoyer/DAFormer.
****

Sound-Guided Semantic Image Manipulation
Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, Sangpil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3377-3386
The recent success of the generative model shows that leveraging the multi-modal embedding space can manipulate an image using text information. However, manipulating an image with other sources rather than text, such as sound, is not easy due to the dynamic characteristics of the sources. Especially, sound can convey vivid emotions and dynamic expressions of the real world. Here, we propose a framework that directly encodes sound into the multi-modal (image-text) embedding space and manipulates an image from the space. Our audio encoder is trained to produce a latent representation from an audio input, which is forced to be aligned with image and text representations in the multi-modal embedding space. We use a direct latent optimization method based on aligned embeddings for sound-guided image manipulation. We also show that our method can mix different modalities, i.e., text and audio, which enrich the variety of the image modification. The experiments on zero-shot audio classification and semantic-level image classificat

ion show that our proposed model outperforms other text and sound-guided state-of-the-art methods.
****

Joint Distribution Matters: Deep Brownian Distance Covariance for Few-Shot Classification
Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, Peihua Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7972-7981
Few-shot classification is a challenging problem as only very few training examples are given for each new task. One of the effective research lines to address this challenge focuses on learning deep representations driven by a similarity measure between a query image and few support images of some class. Statistically, this amounts to measure the dependency of image features, viewed as random vectors in a high-dimensional embedding space. Previous methods either only use marginal distributions without considering joint distributions, suffering from limited representation capability, or are computationally expensive though harnessing joint distributions. In this paper, we propose a deep Brownian Distance Covariance (DeepBDC) method for few-shot classification. The central idea of DeepBDC is to learn image representations by measuring the discrepancy between joint characteristic functions of embedded features and product of the marginals. As the BDC metric is decoupled, we formulate it as a highly modular and efficient layer. Furthermore, we instantiate DeepBDC in two different few-shot classification frameworks. We make experiments on six standard few-shot image benchmarks, covering general object recognition, fine-grained categorization and cross-domain classification. Extensive evaluations show our DeepBDC significantly outperforms the counterparts, while establishing new state-of-the-art results. The source code is available at http://www.peihuali.org/DeepBDC.
****

Proper Reuse of Image Classification Features Improves Object Detection
Cristina Vasconcelos, Vighnesh Birodkar, Vincent Dumoulin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13628-13637
A common practice in transfer learning is to initialize the downstream model weights by pre-training on a data-abundant upstream task. In object detection specifically,the feature backbone is typically initialized with ImageNet classifier weights and fine-tuned on the object detection task. Recent works show this is not strictly necessary under longer training regimes and provide recipes for training the backbone from scratch. We investigate the opposite direction of this end-to-end training trend: we show that an extreme form of knowledge preservation--freezing the classifier-initialized backbone-- consistently improves many different detection models, and leads to considerable resource savings. We hypothesize and corroborate experimentally that the remaining detector components capacity and structure is a crucial factor in leveraging the frozen backbone. Immediate applications of our findings include performance improvements on hard cases like detection of long-tail object classes and computational and memory re-source savings that contribute to making the field more accessible to researchers with access to fewer computational resources.
****

MetaPose: Fast 3D Pose From Multiple Views Without 3D Supervision
Ben Usman, Andrea Tagliasacchi, Kate Saenko, Avneesh Sud; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6759-6770
In the era of deep learning, human pose estimation from multiple cameras with unknown calibration has received little attention to date. We show how to train a neural model to perform this task with high precision and minimal latency overhead. The proposed model takes into account joint location uncertainty due to occlusion from multiple views, and requires only 2D keypoint data for training. Our method outperforms both classical bundle adjustment and weakly-supervised monocular 3D baselines on the well-established Human3.6M dataset, as well as the more challenging in-the-wild Ski-Pose PTZ dataset.

****
End-to-End Human-Gaze-Target Detection With Transformers
Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, Wei Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2202-2210
In this paper, we propose an effective and efficient method for Human-Gaze-Target (HGT) detection, i.e., gaze following. Current approaches decouple the HGT detection task into separate branches of salient object detection and human gaze prediction, employing a two-stage framework where human head locations must first be detected and then be fed into the next gaze target prediction sub-network. In contrast, we redefine the HGT detection task as detecting human head locations and their gaze targets, simultaneously. By this way, our method, named Human-Gaze-Target detection TRansformer or HGTTR, streamlines the HGT detection pipeline by eliminating all other additional components. HGTTR reasons about the relations of salient objects and human gaze from the global image context. Moreover, unlike existing two-stage methods that require human head locations as input and can predict only one human's gaze target at a time, HGTTR can directly predict the locations of all people and their gaze targets at one time in an end-to-end manner. The effectiveness and robustness of our proposed method are verified with extensive experiments on the two standard benchmark datasets, GazeFollowing and VideoAttentionTarget. Without bells and whistles, HGTTR outperforms existing state-of-the-art methods by large margins (6.4 mAP gain on GazeFollowing and 10.3 mAP gain on VideoAttentionTarget) with a much simpler architecture.
****
The Devil Is in the Pose: Ambiguity-Free 3D Rotation-Invariant Learning via Pose-Aware Convolution
Ronghan Chen, Yang Cong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7472-7481
Recent progress in introducing rotation invariance (RI) to 3D deep learning methods is mainly made by designing RI features to replace 3D coordinates as input. The key to this strategy lies in how to restore the global information that is lost by the input RI features. Most state-of-the-arts achieve this by incurring additional blocks or complex global representations, which is time-consuming and ineffective. In this paper, we real that the global information loss stems from an unexplored pose information loss problem, i.e., common convolution layers cannot capture the relative poses between RI features, thus hindering the global information to be hierarchically aggregated in the deep networks. To address this problem, we develop a Pose-aware Rotation Invariant Convolution (i.e., PaRI-Conv), which dynamically adapts its kernels based on the relative poses. Specifically, in each PaRI-Conv layer, a lightweight Augmented Point Pair Feature (APPF) is designed to fully encode the RI relative pose information. Then, we propose to synthesize a factorized dynamic kernel, which reduces the computational cost and memory burden by decomposing it into a shared basis matrix and a pose-aware diagonal matrix that can be learned from the APPF. Extensive experiments on shape classification and part segmentation tasks show that our PaRI-Conv surpasses the state-of-the-art RI methods while being more compact and efficient.
****
Compositional Temporal Grounding With Structured Variational Cross-Graph Correspondence Learning
Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, Xin Eric Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3032-3041
Temporal grounding in videos aims to localize one target video segment that semantically corresponds to a given query sentence. Thanks to the semantic diversity of natural language descriptions, temporal grounding allows activity grounding beyond pre-defined classes and has received increasing attention in recent years. The semantic diversity is rooted in the principle of compositionality in linguistics, where novel semantics can be systematically described by combining known words in novel ways (compositional generalization). However, current temporal grounding datasets do not specifically test for the compositional generalizabilit

y. To systematically measure the compositional generalizability of temporal grou nding models, we introduce a new Compositional Temporal Grounding task and const ruct two new dataset splits, i.e., Charades-CG and ActivityNet-CG. Evaluating th e state-of-the-art methods on our new dataset splits, we empirically find that t hey fail to generalize to queries with novel combinations of seen words. To tack le this challenge, we propose a variational cross-graph reasoning framework that explicitly decomposes video and language into multiple structured hierarchies a nd learns fine-grained semantic correspondence among them. Experiments illustrat e the superior compositional generalizability of our approach.
****

Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline
Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, Xiang Ruan; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8 886-8895
With the popularity of multi-modal sensors, visible-thermal (RGB-T) object track ing is to achieve robust performance and wider application scenarios with the gu idance of objects' temperature information. However, the lack of paired training samples is the main bottleneck for unlocking the power of RGB-T tracking. Since it is laborious to collect high-quality RGB-T sequences, recent benchmarks only provide test sequences. In this paper, we construct a large-scale benchmark wit h high diversity for visible-thermal UAV tracking (VTUAV), including 500 sequenc es with 1.7 million high-resolution (1920*1080 pixels) frame pairs. In addition, comprehensive applications (short-term tracking, long-term tracking and segment ation mask prediction) with diverse categories and scenes are considered for exh austive evaluation. Moreover, we provide a coarse-to-fine attribute annotation, where frame-level attributes are provided to exploit the potential of challenge- specific trackers. In addition, we design a new RGB-T baseline, named Hierarchic al Multi-modal Fusion Tracker (HMFT), which fuses RGB-T data in various levels. Numerous experiments on several datasets are conducted to reveal the effectivene ss of HMFT and the complement of different fusion types. The project is availabl e at here.
****

Future Transformer for Long-Term Action Anticipation
Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 3052-3061
The task of predicting future actions from a video is crucial for a real-world a gent interacting with others. When anticipating actions in the distant future, w e humans typically consider long-term relations over the whole sequence of actio ns, i.e., not only observed actions in the past but also potential actions in th e future. In a similar spirit, we propose an end-to-end attention model for acti on anticipation, dubbed Future Transformer (FUTR), that leverages global attenti on over all input frames and output tokens to predict a minutes-long sequence of future actions. Unlike the previous autoregressive models, the proposed method learns to predict the whole sequence of future actions in parallel decoding, ena bling more accurate and fast inference for long-term anticipation. We evaluate o ur methods on two standard benchmarks for long-term action anticipation, Breakfa st and 50 Salads, achieving state-of-the-art results.
****

Optimal LED Spectral Multiplexing for NIR2RGB Translation
Lei Liu, Yuze Chen, Junchi Yan, Yinqiang Zheng; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12652-12660
The industry practice for night video surveillance is to use auxiliary near-infr ared (NIR) LED diodes, usually centered at 850nm or 940nm, for scene illuminatio n. NIR LED diodes are used to save power consumption while hiding the surveillan ce coverage area from naked human eyes. The captured images are almost monochrom atic, and visual color and texture tend to disappear, which hinders human and ma chine perception. A few existing studies have tried to convert such NIR images t o RGB images through deep learning, which can not provide satisfying results, no r generalize well beyond the training dataset. In this paper, we aim to break th

e fundamental restrictions on reliable NIR-to-RGB (NIR2RGB) translation by exami ning the imaging mechanism of single-chip silicon-based RGB cameras under NIR il luminations, and propose to retrieve the optimal LED multiplexing via deep learn ing. Experimental results show that this translation task can be significantly i mproved by properly multiplexing NIR LEDs close to the visible spectral range th an using 850nm and 940nm LEDs.
****

Rethinking Spatial Invariance of Convolutional Networks for Object Counting
Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, Alexander G. Hauptmann; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19638-19648
Previous work generally believes that improving the spatial invariance of convol utional networks is the key to object counting. However, after verifying several mainstream counting networks, we surprisingly found too strict pixel-level spat ial invariance would cause overfit noise in the density map generation. In this paper, we try to use locally connected Gaussian kernels to replace the original convolution filter to estimate the spatial position in the density map. The purp ose of this is to allow the feature extraction process to potentially stimulate the density map generation process to overcome the annotation noise. Inspired by previous work, we propose a low-rank approximation accompanied with translation invariance to favorably implement the approximation of massive Gaussian convolu tion. Our work points a new direction for follow-up research, which should inves tigate how to properly relax the overly strict pixel-level spatial invariance fo r object counting. We evaluate our methods on 4 mainstream object counting netwo rks (i.e., MCNN, CSRNet, SANet, and ResNet-50). Extensive experiments were condu cted on 7 popular benchmarks for 3 applications (i.e., crowd, vehicle, and plant counting). Experimental results show that our methods significantly outperform other state-of-the-art methods and achieve promising learning of the spatial pos ition of objects.
****

Self-Supervised Video Transformer
Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, Michael S . Ryoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2022, pp. 2874-2884
In this paper, we propose self-supervised training for video transformers using unlabeled video data. From a given video, we create local and global spatiotempo ral views with varying spatial sizes and frame rates. Our self-supervised object ive seeks to match the features of these different views representing the same v ideo, to be invariant to spatiotemporal variations in actions. To the best of ou r knowledge, the proposed approach is the first to alleviate the dependency on n egative samples or dedicated memory banks in Self-supervised Video Transformer (SVT). Further, owing to the flexibility of Transformer models, SVT supports slow -fast video processing within a single architecture using dynamically adjusted p ositional encoding and supports long-term relationship modeling along spatiotemp oral dimensions. Our approach performs well on four action recognition benchmark s (Kinetics-400, UCF-101, HMDB-51, and SSv2) and converges faster with small bat ch sizes. Code is available at: https://git.io/J1juJ
****

AutoRF: Learning 3D Object Radiance Fields From Single View Observations
Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießn er, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 3971-3980
We introduce AutoRF - a new approach for learning neural 3D object representatio ns where each object in the training set is observed by only a single view. This setting is in stark contrast to the majority of existing works that leverage mu ltiple views of the same object, employ explicit priors during training, or requ ire pixel-perfect annotations. To address this challenging setting, we propose t o learn a normalized, object-centric representation whose embedding describes an d disentangles shape, appearance, and pose. Each encoding provides well-generali zable, compact information about the object of interest, which is decoded in a s

ingle-shot into a new target view, thus enabling novel view synthesis. We furthe r improve the reconstruction quality by optimizing shape and appearance codes at test time by fitting the representation tightly to the input image. In a series of experiments, we show that our method generalizes well to unseen objects, eve n across different datasets of challenging real-world street scenes such as nuSc enes, KITTI, and Mapillary Metropolis. Additional results can be found on our pr oject page https://sirwyver.github.io/AutoRF/.
****

Expanding Large Pre-Trained Unimodal Models With Multimodal Information Injectio n for Image-Text Multimodal Classification
Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, Fengmao Lv; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2022, pp. 15492-15501
Fine-tuning pre-trained models for downstream tasks is mainstream in deep learni ng. However, the pre-trained models are limited to be fine-tuned by data from a specific modality. For example, as a visual model, DenseNet cannot directly take the textual data as its input. Hence, although the large pre-trained models suc h as DenseNet or BERT have a great potential for the downstream recognition task s, they have weaknesses in leveraging multimodal information, which is a new tre nd of deep learning. This work focuses on fine-tuning pre-trained unimodal model s with multimodal inputs of image-text pairs and expanding them for image-text m ultimodal recognition. To this end, we propose the Multimodal Information Inject ion Plug-in (MI2P) which is attached to different layers of the unimodal models (e.g., DenseNet and BERT). The proposed MI2P unit provides the path to integrate the information of other modalities into the unimodal models. Specifically, MI2 P performs cross-modal feature transformation by learning the fine-grained corre lations between the visual and textual features. Through the proposed MI2P unit, we can inject the language information into the vision backbone by attending th e word-wise textual features to different visual channels, as well as inject the visual information into the language backbone by attending the channel-wise vis ual features to different textual words. Armed with the MI2P attachments, the pr e-trained unimodal models can be expanded to process multimodal data without the need to change the network structures.
****

Neural RGB-D Surface Reconstruction
Dejan Azinovi■, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, Justus Thies; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2022, pp. 6290-6301
Obtaining high-quality 3D reconstructions of room-scale scenes is of paramount i mportance for upcoming applications in AR or VR. These range from mixed reality applications for teleconferencing, virtual measuring, virtual room planing, to r obotic applications. While current volume-based view synthesis methods that use neural radiance fields (NeRFs) show promising results in reproducing the appeara nce of an object or scene, they do not reconstruct an actual surface. The volume tric representation of the surface based on densities leads to artifacts when a surface is extracted using Marching Cubes, since during optimization, densities are accumulated along the ray and are not used at a single sample point in isola tion. Instead of this volumetric representation of the surface, we propose to re present the surface using an implicit function (truncated signed distance functi on). We show how to incorporate this representation in the NeRF framework, and e xtend it to use depth measurements from a commodity RGB-D sensor, such as a Kine ct. In addition, we propose a pose and camera refinement technique which improve s the overall reconstruction quality. In contrast to concurrent work on integrat ing depth priors in NeRF which concentrates on novel view synthesis, our approac h is able to reconstruct high-quality, metrical 3D reconstructions.
****

ClusterGNN: Cluster-Based Coarse-To-Fine Graph Neural Network for Efficient Feat ure Matching
Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, Kai Zhang; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV

PR), 2022, pp. 12517-12526
Graph Neural Networks (GNNs) with attention have been successfully applied for learning visual feature matching. However, current methods learn with complete graphs, resulting in a quadratic complexity in the number of features. Motivated by a prior observation that self- and cross- attention matrices converge to a sparse representation, we propose ClusterGNN, an attentional GNN architecture which operates on clusters for learning the feature matching task. Using a progressive clustering module we adaptively divide keypoints into different subgraphs to reduce redundant connectivity, and employ a coarse-to-fine paradigm for mitigating miss-classification within images. Our approach yields a 59.7% reduction in runtime and 58.4% reduction in memory consumption for dense detection, compared to current state-of-the-art GNN-based matching, while achieving a competitive performance on various computer vision tasks.
****
AdaptPose: Cross-Dataset Adaptation for 3D Human Pose Estimation by Learnable Motion Generation
Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, Z. Jane Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13075-13085
This paper addresses the problem of cross-dataset generalization of 3D human pose estimation models. Testing a pre-trained 3D pose estimator on a new dataset results in a major performance drop. Previous methods have mainly addressed this problem by improving the diversity of the training data. We argue that diversity alone is not sufficient and that the characteristics of the training data need to be adapted to those of the new dataset such as camera viewpoint, position, human actions, and body size. To this end, we propose AdaptPose, an end-to-end framework that generates synthetic 3D human motions from a source dataset and uses them to fine-tune a 3D pose estimator. AdaptPose follows an adversarial training scheme. From a source 3D pose the generator generates a sequence of 3D poses and a camera orientation that is used to project the generated poses to a novel view. Without any 3D labels or camera information AdaptPose successfully learns to create synthetic 3D poses from the target dataset while only being trained on 2D poses. In experiments on the Human3.6M, MPI-INF-3DHP, 3DPW, and Ski-Pose datasets our method outperforms previous work in cross-dataset evaluations by 14% and previous semi-supervised learning methods that use partial 3D annotations by 16%.
****
ClothFormer: Taming Video Virtual Try-On in All Module
Jianbin Jiang, Tan Wang, He Yan, Junhui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10799-10808
The task of video virtual try-on aims to fit the target clothes to a person in the video with spatio-temporal consistency. Despite tremendous progress of image virtual try-on, they lead to inconsistency between frames when applied to videos. Limited work also explored the task of video-based virtual try-on but failed to produce visually pleasing and temporally coherent results. Moreover, there are two other key challenges: 1) how to generate accurate warping when occlusions appear in the clothing region; 2) how to generate clothes and non-target body parts (e.g. arms, neck) in harmony with the complicated background; To address them, we propose a novel video virtual try-on framework, ClothFormer, which successfully synthesizes realistic, harmonious, and spatio-temporal consistent results in complicated environment. In particular, ClothFormer involves three major modules. First, a two-stage anti-occlusion warping module that predicts an accurate dense flow mapping between the body regions and the clothing regions. Second, an appearance-flow tracking module utilizes ridge regression and optical flow correction to smooth the dense flow sequence and generate a temporally smooth warped clothing sequence. Third, a dual-stream transformer extracts and fuses clothing textures, person features, and environment information to generate realistic try-on videos. Through rigorous experiments, we demonstrate that our method highly surpasses the baselines in terms of synthesized video quality both qualitatively and quantitatively.

****
Cross-Domain Adaptive Teacher for Object Detection
Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijia n He, Kris Kitani, Peter Vajda; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2022, pp. 7581-7590
We address the task of domain adaptation in object detection, where there is a d omain gap between a domain with annotations (source) and a domain of interest wi thout annotations (target). As an effective semi-supervised learning method, the teacher-student framework (a student model is supervised by the pseudo labels f rom a teacher model) has also yielded a large accuracy gain in cross-domain obje ct detection. However, it suffers from the domain shift and generates many low-q uality pseudo labels (e.g., false positives), which leads to sub-optimal perform ance. To mitigate this problem, we propose a teacher-student framework named Ada ptive Teacher (AT) which leverages domain adversarial learning and weak-strong d ata augmentation to address the domain gap. Specifically, we employ feature-leve l adversarial training in the student model, allowing features derived from the source and target domains to share similar distributions. This process ensures t he student model produces domain-invariant features. Furthermore, we apply weak-strong augmentation and mutual learning between the teacher model (taking data f rom the target domain) and the student model (taking data from both domains). Th is enables the teacher model to learn the knowledge from the student model witho ut being biased to the source domain. We show that AT demonstrates superiority o ver existing approaches and even Oracle (fully-supervised) models by a large mar gin. For example, we achieve 50.9% (49.3%) mAP on Foggy Cityscape (Clipart1K), w hich is 9.2% (5.2%) and 8.2% (11.0%) higher than previous state-of-the-art and O racle, respectively.
****
Geometric Anchor Correspondence Mining With Uncertainty Modeling for Universal D omain Adaptation
Liang Chen, Yihang Lou, Jianzhong He, Tao Bai, Minghua Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16134-16143
Universal domain adaptation (UniDA) aims to transfer the knowledge learned from a label-rich source domain to a label-scarce target domain without any constrain ts on the label space. However, domain shift and category shift make UniDA extre mely challenging, which mainly lies in how to recognize both shared "known" samp les and private "unknown" samples. Previous works rarely explore the intrinsic g eometrical relationship between the two domains, and they manually set a thresho ld for the overconfident closed-world classifier to reject "unknown" samples. Th erefore, in this paper, we propose a Geometric anchor-guided Adversarial and con Trastive learning framework with uncErtainty modeling called GATE to alleviate t hese issues. Specifically, we first develop a random walk-based anchor mining st rategy together with a high-order attention mechanism to build correspondence ac ross domains. Then a global joint local domain alignment paradigm is designed, i .e., geometric adversarial learning for global distribution calibration and subg raph-level contrastive learning for local region aggregation. Toward accurate ta rget private samples detection, GATE introduces a universal incremental classifi er by modeling the energy uncertainty. We further efficiently generate novel cat egories by manifold mixup, and minimize the open-set entropy to learn the "unkno wn" threshold adaptively. Extensive experiments on three benchmarks demonstrate that GATE significantly outperforms previous state-of-the-art UniDA methods.
****
Class-Balanced Pixel-Level Self-Labeling for Domain Adaptive Semantic Segmentati on
Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 22, pp. 11593-11603
Domain adaptive semantic segmentation aims to learn a model with the supervision of source domain data, and produce satisfactory dense predictions on unlabeled target domain. One popular solution to this challenging task is self-training, w

hich selects high-scoring predictions on target samples as pseudo labels for tra ining. However, the produced pseudo labels often contain much noise because the model is biased to source domain as well as majority categories. To address the above issues, we propose to directly explore the intrinsic pixel distributions o f target domain data, instead of heavily relying on the source domain. Specifica lly, we simultaneously cluster pixels and rectify pseudo labels with the obtaine d cluster assignments. This process is done in an online fashion so that pseudo labels could co-evolve with the segmentation model without extra training rounds . To overcome the class imbalance problem on long-tailed categories, we employ a distribution alignment technique to enforce the marginal class distribution of cluster assignments to be close to that of pseudo labels. The proposed method, n amely Class-balanced Pixel-level Self-Labeling (CPSL), improves the segmentation performance on target domain over state-of-the-arts by a large margin, especial ly on long-tailed categories.

****

Coopernaut: End-to-End Driving With Cooperative Perception for Networked Vehicle s

Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, Yuke Zhu; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1725 2-17262

Optical sensors and learning algorithms for autonomous vehicles have dramaticall y advanced in the past few years. Nonetheless, the reliability of today's autono mous vehicles is hindered by the limited line-of-sight sensing capability and th e brittleness of data-driven methods in handling extreme situations. With recent developments of telecommunication technologies, cooperative perception with veh icle-to-vehicle communications has become a promising paradigm to enhance autono mous driving in dangerous or emergency situations. We introduce COOPERNAUT, an e nd-to-end learning model that uses cross-vehicle perception for vision-based coo perative driving. Our model encodes LiDAR information into compact point-based r epresentations that can be transmitted as messages between vehicles via realisti c wireless channels. To evaluate our model, we develop AutoCastSim, a network-au gmented driving simulation framework with example accident-prone scenarios. Our experiments on AutoCastSim suggest that our cooperative perception driving model s lead to a 40% improvement in average success rate over egocentric driving mode ls in these challenging driving situations and a 5 times smaller bandwidth requi rement than prior work V2VNet. COOPERNAUT and AUTOCASTSIM are available at https ://ut-austin-rpl.github.io/Coopernaut/.

****

Condensing CNNs With Partial Differential Equations

Anil Kag, Venkatesh Saligrama; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2022, pp. 610-619

Convolutional neural networks (CNNs) rely on the depth of the architecture to ob tain complex features. It results in computationally expensive models for low-re source IoT devices. Convolutional operators are local and restricted in the rece ptive field, which increases with depth. We explore partial differential equatio ns (PDEs) that offer a global receptive field without the added overhead of main taining large kernel convolutional filters. We propose a new feature layer, call ed the Global layer, that enforces PDE constraints on the feature maps, resultin g in rich features. These constraints are solved by embedding iterative schemes in the network. The proposed layer can be embedded in any deep CNN to transform it into a shallower network. Thus, resulting in compact and computationally effi cient architectures achieving similar performance as the original network. Our e xperimental evaluation demonstrates that architectures with global layers requir e 2-5xless computational and storage budget without any significant loss in perf ormance.

****

Few-Shot Keypoint Detection With Uncertainty Learning for Unseen Species

Changsheng Lu, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19416-19426

Current non-rigid object keypoint detectors perform well on a chosen kind of spe

cies and body parts, and require a large amount of labelled keypoints for traini
ng. Moreover, their heatmaps, tailored to specific body parts, cannot recognize
 novel keypoints (keypoints not labelled for training) on unseen species. We rais
e an interesting yet challenging question: how to detect both base (annotated fo
r training) and novel keypoints for unseen species given a few annotated samples
? Thus, we propose a versatile Few-shot Keypoint Detection (FSKD) pipeline, whic
h can detect a varying number of keypoints of different kinds. Our FSKD provides
 the uncertainty estimation of predicted keypoints. Specifically, FSKD involves
main and auxiliary keypoint representation learning, similarity learning, and ke
ypoint localization with uncertainty modeling to tackle the localization noise.
Moreover, we model the uncertainty across groups of keypoints by multivariate Ga
ussian distribution to exploit implicit correlations between neighboring keypoin
ts. We show the effectiveness of our FSKD on (i) novel keypoint detection for un
seen species, and (ii) few-shot Fine-Grained Visual Recognition (FGVR) and (iii)
 Semantic Alignment (SA) downstream tasks. For FGVR, detected keypoints improve
the classification accuracy. For SA, we showcase a novel thin-plate-spline warpi
ng that uses estimated keypoint uncertainty under imperfect keypoint coresponden
ces.
****

Improving Robustness Against Stealthy Weight Bit-Flip Attacks by Output Code Mat
ching

Ozan Özdenizci, Robert Legenstein; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2022, pp. 13388-13397

Deep neural networks (DNNs) have been shown to be vulnerable against adversarial
 weight bit-flip attacks through hardware-induced fault-injection methods on the
 memory systems where network parameters are stored. Recent attacks pose the fur
ther concerning threat of finding minimal targeted and stealthy weight bit-flips
 that preserve expected behavior for untargeted test samples. This renders the a
ttack undetectable from a DNN operation perspective. We propose a DNN defense me
chanism to improve robustness in such realistic stealthy weight bit-flip attack
scenarios. Our output code matching networks use an output coding scheme where t
he usual one-hot encoding of classes is replaced by partially overlapping bit st
rings. We show that this encoding significantly reduces attack stealthiness. Imp
ortantly, our approach is compatible with existing defenses and DNN architecture
s. It can be efficiently implemented on pre-trained models by simply re-defining
 the output classification layer and finetuning. Experimental benchmark evaluati
ons show that output code matching is superior to existing regularized weight qu
antization based defenses, and an effective defense against stealthy weight bit-
flip attacks.
****

Unsupervised Hierarchical Semantic Segmentation With Multiview Cosegmentation an
d Clustering Transformers

Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, Stella X. Yu; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
022, pp. 2571-2581

Unsupervised semantic segmentation aims to discover groupings within and across
images that capture object- and view-invariance of a category without external s
upervision. Grouping naturally has levels of granularity, creating ambiguity in
unsupervised segmentation. Existing methods avoid this ambiguity and treat it as
 a factor outside modeling, whereas we embrace it and desire hierarchical groupi
ng consistency for unsupervised segmentation. We approach unsupervised segmentat
ion as a pixel-wise feature learning problem. Our idea is that a good representa
tion must be able to reveal not just a particular level of grouping, but any lev
el of grouping in a consistent and predictable manner across different levels of
 granularity. We enforce spatial consistency of grouping and bootstrap feature l
earning with co-segmentation among multiple views of the same image, and enforce
 semantic consistency across the grouping hierarchy with clustering transformers
. We deliver the first data-driven unsupervised hierarchical semantic segmentati
on method called Hierarchical Segment Grouping (HSG). Capturing visual similarit
y and statistical co-occurrences, HSG also outperforms existing unsupervised seg

mentation methods by a large margin on five major object- and scene-centric benc
hmarks.
****

3D-SPS: Single-Stage 3D Visual Grounding via Referred Point Progressive Selectio
n

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia
, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2022, pp. 16454-16463

3D visual grounding aims to locate the referred target object in 3D point cloud
scenes according to a free-form language description. Previous methods mostly fo
llow a two-stage paradigm, i.e., language-irrelevant detection and cross-modal m
atching, which is limited by the isolated architecture. In such a paradigm, the
detector needs to sample keypoints from raw point clouds due to the inherent pro
perties of 3D point clouds (irregular and large-scale), to generate the correspo
nding object proposal for each keypoint. However, sparse proposals may leave out
 the target in detection, while dense proposals may confuse the matching model.
Moreover, the language-irrelevant detection stage can only sample a small propor
tion of keypoints on the target, deteriorating the target prediction. In this pa
per, we propose a 3D Single-Stage Referred Point Progressive Selection (3D-SPS)
method, which progressively selects keypoints with the guidance of language and
directly locates the target. Specifically, we propose a Description-aware Keypoi
nt Sampling (DKS) module to coarsely focus on the points of language-relevant ob
jects, which are significant clues for grounding. Besides, we devise a Target-or
iented Progressive Mining (TPM) module to finely concentrate on the points of th
e target, which is enabled by progressive intra-modal relation modeling and inte
r-modal target mining. 3D-SPS bridges the gap between detection and matching in
the 3D visual grounding task, localizing the target at a single stage. Experimen
ts demonstrate that 3D-SPS achieves state-of-the-art performance on both ScanRef
er and Nr3D/Sr3D datasets.
****

TubeR: Tubelet Transformer for Video Action Detection

Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui L
iu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek,
Joseph Tighe; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2022, pp. 13598-13607

We propose TubeR: a simple solution for spatio-temporal video action detection.
Different from existing methods that depend on either an off-line actor detector
 or hand-designed actor-positional hypotheses like proposals or anchors, we prop
ose to directly detect an action tubelet in video by simultaneously performing a
ction localization and recognition from a single representation. TubeR learns a
set of tubelet-queries and utilizes a tubelet-attention module to model the dyna
mic spatio-temporal nature of a video clip, which effectively reinforces the mod
el capacity compared to using actor-positional hypotheses in the spatio-temporal
 space. For videos containing transitional states or scene changes, we propose a
 context aware classification head to utilize short-term and long-term context t
o strengthen action classification, and an action switch regression head for det
ecting the precise temporal action extent. TubeR directly produces action tubele
ts with variable lengths and even maintains good results for long video clips. T
ubeR outperforms the previous state-of-the-art on commonly used action detection
 datasets AVA, UCF101-24 and JHMDB51-21. Code will be available on GluonCV(https
://cv.gluon.ai/).
****

LASER: LAtent SpacE Rendering for 2D Visual Localization

Zhixiang Min, Naji Khosravan, Zachary Bessinger, Manjunath Narayana, Sing Bing K
ang, Enrique Dunn, Ivaylo Boyadzhiev; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11122-11131

We present LASER, an image-based Monte Carlo Localization (MCL) framework for 2D
 floor maps. LASER introduces the concept of latent space rendering, where 2D po
se hypotheses on the floor map are directly rendered into a geometrically-struct
ured latent space by aggregating viewing ray features. Through a tightly coupled

rendering codebook scheme, the viewing ray features are dynamically determined at rendering-time based on their geometries (i.e. length, incident-angle), endowing our representation with view-dependent fine-grain variability. Our codebook scheme effectively disentangles feature encoding from rendering, allowing the latent space rendering to run at speeds above 10KHz. Moreover, through metric learning, our geometrically-structured latent space is common to both pose hypotheses and query images with arbitrary field of views. As a result, LASER achieves state-of-the-art performance on large-scale indoor localization datasets (i.e. ZInD and Structured3D) for both panorama and perspective image queries, while significantly outperforming existing learning-based methods in speed.

****

MUM: Mix Image Tiles and UnMix Feature Tiles for Semi-Supervised Object Detection

JongMok Kim, JooYoung Jang, Seunghyeon Seo, Jisoo Jeong, Jongkeun Na, Nojun Kwak ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14512-14521

Many recent semi-supervised learning (SSL) studies build teacher-student architecture and train the student network by the generated supervisory signal from the teacher. Data augmentation strategy plays a significant role in the SSL framework since it is hard to create a weak-strong augmented input pair without losing label information. Especially when extending SSL to semi-supervised object detection (SSOD), many strong augmentation methodologies related to image geometry and interpolation-regularization are hard to utilize since they possibly hurt the location information of the bounding box in the object detection task. To address this, we introduce a simple yet effective data augmentation method, Mix/UnMix (MUM), which unmixes feature tiles for the mixed image tiles for the SSOD framework. Our proposed method makes mixed input image tiles and reconstructs them in the feature space. Thus, MUM can enjoy the interpolation-regularization effect from non-interpolated pseudo-labels and successfully generate a meaningful weak-strong pair. Furthermore, MUM can be easily equipped on top of various SSOD methods. Extensive experiments on MS-COCO and PASCAL VOC datasets demonstrate the superiority of MUM by consistently improving the mAP performance over the baseline in all the tested SSOD benchmark protocols. The code is released at https://github.com/JongMokKim/mix-unmix.

****

On Adversarial Robustness of Trajectory Prediction for Autonomous Vehicles

Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, Z. Morley Mao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15159-15168

Trajectory prediction is a critical component for autonomous vehicles (AVs) to perform safe planning and navigation. However, few studies have analyzed the adversarial robustness of trajectory prediction or investigated whether the worst-case prediction can still lead to safe planning. To bridge this gap, we study the adversarial robustness of trajectory prediction models by proposing a new adversarial attack that perturbs normal vehicle trajectories to maximize the prediction error. Our experiments on three models and three datasets show that the adversarial prediction increases the prediction error by more than 150%. Our case studies show that if an adversary drives a vehicle close to the target AV following the adversarial trajectory, the AV may make an inaccurate prediction and even make unsafe driving decisions. We also explore possible mitigation techniques via data augmentation and trajectory smoothing.

****

Kubric: A Scalable Dataset Generator

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi , Fangcheng Zhong, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference o

n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3749-3761
Data is the driving force of machine learning, with the amount and quality of tr aining data often being more important for the performance of a system than arch itecture and training details. But collecting, processing and annotating real da ta at scale is difficult, expensive, and frequently raises additional privacy, f airness and legal concerns. Synthetic data is a powerful tool with the potential to address these shortcomings: 1) it is cheap 2) supports rich ground-truth ann otations 3) offers full control over data and 4) can circumvent or mitigate prob lems regarding bias, privacy and licensing. Unfortunately, software tools for ef fective data generation are less mature than those for architecture design and t raining, which leads to fragmented generation efforts. To address these problems we introduce Kubric, an open-source Python framework that interfaces with PyBul let and Blender to generate photo-realistic scenes, with rich annotations, and s eamlessly scales to large jobs distributed over thousands of machines, and gener ating TBs of data. We demonstrate the effectiveness of Kubric by presenting a se ries of 11 different generated datasets for tasks ranging from studying 3D NeRF models to optical flow estimation. We release Kubric, the used assets, all of th e generation code, as well as the rendered datasets for reuse and modification.
****

Unpaired Deep Image Deraining Using Dual Contrastive Learning
Xiang Chen, Jinshan Pan, Kui Jiang, Yufeng Li, Yufeng Huang, Caihua Kong, Longga ng Dai, Zhentao Fan; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2022, pp. 2017-2026
Learning single image deraining (SID) networks from an unpaired set of clean and rainy images is practical and valuable as acquiring paired real-world data is a lmost infeasible. However, without the paired data as the supervision, learning a SID network is challenging. Moreover, simply using existing unpaired learning methods (e.g., unpaired adversarial learning and cycle-consistency constraints) in the SID task is insufficient to learn the underlying relationship from rainy inputs to clean outputs as there exists significant domain gap between the rainy and clean images. In this paper, we develop an effective unpaired SID adversari al framework which explores mutual properties of the unpaired exemplars by a dua l contrastive learning manner in a deep feature space, named as DCD-GAN. The pro posed method mainly consists of two cooperative branches: Bidirectional Translat ion Branch (BTB) and Contrastive Guidance Branch (CGB). Specifically, BTB exploi ts full advantage of the circulatory architecture of adversarial consistency to generate abundant exemplar pairs and excavates latent feature distributions betw een two domains by equipping it with bidirectional mapping. Simultaneously, CGB implicitly constrains the embeddings of different exemplars in the deep feature space by encouraging the similar feature distributions closer while pushing the dissimilar further away, in order to better facilitate rain removal and help ima ge restoration. Extensive experiments demonstrate that our method performs favor ably against existing unpaired deraining approaches on both synthetic and real-w orld datasets, and generates comparable results against several fully-supervised or semi-supervised models.
****

Learning Multiple Dense Prediction Tasks From Partially Annotated Data
Wei-Hong Li, Xialei Liu, Hakan Bilen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18879-18889
Despite the recent advances in multi-task learning of dense prediction problems, most methods rely on expensive labelled datasets. In this paper, we present a l abel efficient approach and look at jointly learning of multiple dense predictio n tasks on partially annotated data (i.e. not all the task labels are available for each image), which we call multi-task partially-supervised learning. We prop ose a multi-task training procedure that successfully leverages task relations t o supervise its multi-task learning when data is partially annotated. In particu lar, we learn to map each task pair to a joint pairwise task-space which enables sharing information between them in a computationally efficient way through ano ther network conditioned on task pairs, and avoids learning trivial cross-task r elations by retaining high-level information about the input image. We rigorousl

y demonstrate that our proposed method effectively exploits the images with unlabelled tasks and outperforms existing semi-supervised learning approaches and related methods on three standard benchmarks.
****

Pushing the Performance Limit of Scene Text Recognizer Without Human Annotation
Caiyuan Zheng, Hui Li, Seon-Min Rhee, Seungju Han, Jae-Joon Han, Peng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14116-14125
Scene text recognition (STR) attracts much attention over the years because of its wide application. Most methods train STR model in a fully supervised manner which requires large amounts of labeled data. Although synthetic data contributes a lot to STR, it suffers from the real-to-synthetic domain gap that restricts model performance. In this work, we aim to boost STR models by leveraging both synthetic data and the numerous real unlabeled images, exempting human annotation cost thoroughly. A robust consistency regularization based semi-supervised framework is proposed for STR, which can effectively solve the instability issue due to domain inconsistency between synthetic and real images. A character-level consistency regularization is designed to mitigate the misalignment between characters in sequence recognition. Extensive experiments on standard text recognition benchmarks demonstrate the effectiveness of the proposed method. It can steadily improve existing STR models, and boost an STR model to achieve new state-of-the-art results. To our best knowledge, this is the first consistency regularization based framework that applies successfully to STR.
****

Boosting 3D Object Detection by Simulating Multimodality on Point Clouds
Wu Zheng, Mingxuan Hong, Li Jiang, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13638-13647
This paper presents a new approach to boost a single-modality (LiDAR) 3D object detector by teaching it to simulate features and responses that follow a multi-modality (LiDAR-image) detector. The approach needs LiDAR-image data only when training the single-modality detector, and once well-trained, it only needs LiDAR data at inference. We design a novel framework to realize the approach: response distillation to focus on the crucial response samples and avoid the background samples; sparse-voxel distillation to learn voxel semantics and relations from the estimated crucial voxels; a fine-grained voxel-to-point distillation to better attend to features of small and distant objects; and instance distillation to further enhance the deep-feature consistency. Experimental results on the nuScenes dataset show that our approach outperforms all SOTA LiDAR-only 3D detectors and even surpasses the baseline LiDAR-image detector on the key NDS metric, filling 72% mAP gap between the single- and multi-modality detectors.
****

Towards Low-Cost and Efficient Malaria Detection
Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, Mohsen Ali; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20687-20696
Malaria, a fatal but curable disease claims hundreds of thousands of lives every year. Early and correct diagnosis is vital to avoid health complexities, however, it depends upon the availability of costly microscopes and trained experts to analyze blood-smear slides. Deep learning-based methods have the potential to not only decrease the burden of experts but also improve diagnostic accuracy on low-cost microscopes. However, this is hampered by the absence of a reasonable size dataset. One of the most challenging aspects is the reluctance of the experts to annotate the dataset at low magnification on low-cost microscopes. We present a dataset to further the research on malaria microscopy over the low-cost microscopes at low magnification. Our large-scale dataset consists of images of blood-smear slides from several malaria-infected patients, collected through microscopes at two different cost spectrums and multiple magnifications. Malarial cells are annotated for the localization and life-stage classification task on the images collected through the high-cost microscope at high magnification. We design a mechanism to transfer these annotations from the high-cost microscope at high

magnification to the low-cost microscope, at multiple magnifications. Multiple object detectors and domain adaptation methods are presented as the baselines. Furthermore, a partially supervised domain adaptation method is introduced to adapt the object-detector to work on the images collected from the low-cost microscope. The dataset and benchmark models will be made publicly available.

****

Learning Neural Light Fields With Ray-Space Embedding

Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, Changil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19819-19829

Neural radiance fields (NeRFs) produce state-of-the-art view synthesis results, but are slow to render, requiring hundreds of network evaluations per pixel to approximate a volume rendering integral. Baking NeRFs into explicit data structures enables efficient rendering, but results in large memory footprints and, in some cases, quality reduction. Additionally, volumetric representations for view synthesis often struggle to represent challenging view dependent effects such as distorted reflections and refractions. We present a novel neural light field representation that, in contrast to prior work, is fast, memory efficient, and excels at modeling complicated view dependence. Our method supports rendering with a single network evaluation per pixel for small baseline light fields and with only a few evaluations per pixel for light fields with larger baselines. At the core of our approach is a ray-space embedding network that maps 4D ray-space into an intermediate, interpolable latent space. Our method achieves state-of-the-art quality on dense forward-facing datasets such as the Stanford Light Field dataset. In addition, for forward-facing scenes with sparser inputs we achieve results that are competitive with NeRF-based approaches while providing a better speed/quality/memory trade-off with far fewer network evaluations.

****

Exposure Normalization and Compensation for Multiple-Exposure Correction

Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6043-6052

Images captured with improper exposures usually bring unsatisfactory visual effects. Previous works mainly focus on either underexposure or overexposure correction, resulting in poor generalization to various exposures. An alternative solution is to mix the multiple exposure data for training a single network. However, the procedures of correcting underexposure and overexposure to normal exposures are much different from each other, leading to large discrepancies for the network in correcting multiple exposures, thus resulting in poor performance. The key point to address this issue lies in bridging different exposure representations. To achieve this goal, we design a multiple exposure correction framework based on an Exposure Normalization and Compensation (ENC) module. Specifically, the ENC module consists of an exposure normalization part for mapping different exposure features to the exposure-invariant feature space, and a compensation part for integrating the initial features unprocessed by exposure normalization part to ensure the completeness of information. Besides, to further alleviate the imbalanced performance caused by variations in the optimization process, we introduce a parameter regularization fine-tuning strategy to improve the performance of the worst-performed exposure without degrading other exposures. Our model empowered by ENC outperforms the existing methods by more than 2dB and is robust to multiple image enhancement tasks, demonstrating its effectiveness and generalization capability for real-world applications.

****

UDA-COPE: Unsupervised Domain Adaptation for Category-Level Object Pose Estimation

Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14891-14900

Learning to estimate object pose often requires ground-truth (GT) labels, such as CAD model and absolute-scale object pose, which is expensive and laborious to

obtain in the real world. To tackle this problem, we propose an unsupervised dom ain adaptation (UDA) for category-level object pose estimation, called UDA-COPE. Inspired by recent multi-modal UDA techniques, the proposed method exploits a t eacher-student self-supervised learning scheme to train a pose estimation networ k without using target domain pose labels. We also introduce a bidirectional fil tering method between the predicted normalized object coordinate space (NOCS) ma p and observed point cloud, to not only make our teacher network more robust to the target domain but also to provide more reliable pseudo labels for the studen t network training. Extensive experimental results demonstrate the effectiveness of our proposed method both quantitatively and qualitatively. Notably, without leveraging target-domain GT labels, our proposed method achieved comparable or s ometimes superior performance to existing methods that depend on the GT labels.

****

Learning Non-Target Knowledge for Few-Shot Semantic Segmentation

Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, Ling Shao; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11573-11582

Existing studies in few-shot semantic segmentation only focus on mining the targ et object information, however, often are hard to tell ambiguous regions, especi ally in non-target regions, which include background (BG) and Distracting Object s (DOs). To alleviate this problem, we propose a novel framework, namely Non-Tar get Region Eliminating (NTRE) network, to explicitly mine and eliminate BG and D O regions in the query. First, a BG Mining Module (BGMM) is proposed to extract the BG region via learning a general BG prototype. To this end, we design a BG l oss to supervise the learning of BGMM only using the known target object segment ation ground truth. Then, a BG Eliminating Module and a DO Eliminating Module ar e proposed to successively filter out the BG and DO information from the query f eature, based on which we can obtain a BG and DO-free target object segmentation result. Furthermore, we propose a prototypical contrastive learning algorithm t o improve the model ability of distinguishing the target object from DOs. Extens ive experiments on both PASCAL- 5^i and COCO- 20^i datasets show that our approach is effective despite its simplicity. Code is available at https://githu b.com/LIUYUANWEI98/NERTNet

****

TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection With Transformer s

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, Chiew-Lan Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2022, pp. 1090-1099

LiDAR and camera are two important sensors for 3D object detection in autonomous driving. Despite the increasing popularity of sensor fusion in this field, the robustness against inferior image conditions, e.g., bad illumination and sensor misalignment, is under-explored. Existing fusion methods are easily affected by such conditions, mainly due to a hard association of LiDAR points and image pixe ls, established by calibration matrices. We propose TransFusion, a robust soluti on to LiDAR-camera fusion with a soft-association mechanism to handle inferior i mage conditions. Specifically, our TransFusion consists of convolutional backbon es and a detection head based on a transformer decoder. The first layer of the d ecoder predicts initial bounding boxes from a LiDAR point cloud using a sparse s et of object queries, and its second decoder layer adaptively fuses the object q ueries with useful image features, leveraging both spatial and contextual relati onships. The attention mechanism of the transformer enables our model to adaptiv ely determine where and what information should be taken from the image, leading to a robust and effective fusion strategy. We additionally design an image-guid ed query initialization strategy to deal with objects that are difficult to dete ct in point clouds. TransFusion achieves state-of-the-art performance on large-s cale datasets. We provide extensive experiments to demonstrate its robustness ag ainst degenerated image quality and calibration errors. We also extend the propo sed method to the 3D tracking task and achieve the 1st place in the leaderboard of nuScenes tracking, showing its effectiveness and generalization capability.

****
Real-Time Hyperspectral Imaging in Hardware via Trained Metasurface Encoders
Maksim Makarenko, Arturo Burguete-Lopez, Qizhou Wang, Fedor Getman, Silvio Giancola, Bernard Ghanem, Andrea Fratalocchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12692-12702
Hyperspectral imaging has attracted significant attention to identify spectral signatures for image classification and automated pattern recognition in computer vision. State-of-the-art implementations of snapshot hyperspectral imaging rely on bulky, non-integrated, and expensive optical elements, including lenses, spectrometers, and filters. These macroscopic components do not allow fast data processing for, e.g., real-time and high-resolution videos. This work introduces Hyplex, a new integrated architecture addressing the limitations discussed above. Hyplex is a CMOS-compatible, fast hyperspectral camera that replaces bulk optics with nanoscale metasurfaces inversely designed through artificial intelligence. Hyplex does not require spectrometers but makes use of conventional monochrome cameras, opening up real-time and high-resolution hyperspectral imaging at inexpensive costs. Hyplex exploits a model-driven optimization, which connects the physical metasurfaces layer with modern visual computing approaches based on end-to-end training. We design and implement a prototype version of Hyplex and compare its performance against the state-of-the-art for typical imaging tasks such as spectral reconstruction and semantic segmentation. In all benchmarks, Hyplex reports the smallest reconstruction error. In addition, to the best of the authors' knowledge, we created FVgNET, the largest publicly available labeled hyperspectral dataset for semantic segmentation tasks.
****
Clean Implicit 3D Structure From Noisy 2D STEM Images
Hannah Kniesel, Timo Ropinski, Tim Bergner, Kavitha Shaga Devan, Clarissa Read, Paul Walther, Tobias Ritschel, Pedro Hermosilla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20762-20772
Scanning Transmission Electron Microscopes (STEMs) acquire 2D images of a 3D sample on the scale of individual cell components. Unfortunately, these 2D images can be too noisy to be fused into a useful 3D structure and facilitating good denoisers is challenging due to the lack of clean-noisy pairs. Additionally, representing detailed 3D structure can be difficult even for clean data when using regular 3D grids. Addressing these two limitations, we suggest a differentiable image formation model for STEM, allowing to learn a joint model of 2D sensor noise in STEM together with an implicit 3D model. We show, that the combination of these models are able to successfully disentangle 3D signal and noise without supervision and outperform at the same time several baselines on synthetic and real data.
****
UKPGAN: A General Self-Supervised Keypoint Detector
Yang You, Wenhai Liu, Yanjie Ze, Yong-Lu Li, Weiming Wang, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17042-17051
Keypoint detection is an essential component for the object registration and alignment. In this work, we reckon keypoint detection as information compression, and force the model to distill out important points of an object. Based on this, we propose UKPGAN, a general self-supervised 3D keypoint detector where keypoints are detected so that they could reconstruct the original object shape. Two modules: GAN-based keypoint sparsity control and salient information distillation modules are proposed to locate those important keypoints. Extensive experiments show that our keypoints align well with human annotated keypoint labels, and can be applied to SMPL human bodies under various non-rigid deformations. Furthermore, our keypoint detector trained on clean object collections generalizes well to real-world scenarios, thus further improves geometric registration when combined with off-the-shelf point descriptors. Repeatability experiments show that our model is stable under both rigid and non-rigid transformations, with local reference frame estimation. Our code is available on https://github.com/qq456cvb/UKPGAN.

****
Learning Optimal K-Space Acquisition and Reconstruction Using Physics-Informed N eural Networks
Wei Peng, Li Feng, Guoying Zhao, Fang Liu; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20794-20803
The inherent slow imaging speed of Magnetic Resonance Image (MRI) has spurred th e development of various acceleration methods, typically through heuristically u ndersampling of the associated measurement domain known as k-space. Recently, de ep neural networks have been applied to reconstruct undersampled k-space and sho wn improved reconstruction performance. While most methods focus on designing no vel reconstruction networks or new training strategies for a given undersampling pattern, e.g., random Cartesian undersampling or standard non-Cartesian samplin g, to date, there is limited research that aims to learn and optimize k-space sa mpling strategies using deep neural networks. In this work, we propose a novel f ramework to learn optimized k-space sampling trajectories using deep learning by considering it as an Ordinary Differential Equation (ODE) problem that can be s olved using neural ODE. In particular, the sampling of k-space data is framed as a dynamic system, in which the control points serve as an initial state and a p hysical-conditioned neural ODE is formulated to approximate the system. Moreover , we also enforce additional constraints on gradient slew rate and amplitude in trajectory learning, so that severe gradient-indued artifacts can be minimized. Furthermore, we have also demonstrated that sampling trajectory optimization and MRI reconstruction can be jointly trained, such that the optimized trajectory i s task-oriented and can enhance overall image reconstruction performance. Experi ments were conducted on different in-vivo dataset (e.g., Brain and Knee) with di fferent contrast. Initial results have shown that our proposed method is able to generate better image quality in accelerated MRI compared to conventional under sampling schemes in both Cartesian and non-Cartesian acquisitions.
****
Leveraging Adversarial Examples To Quantify Membership Information Leakage
Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, Pablo Pi antanida; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10399-10409
The use of personal data for training machine learning systems comes with a priv acy threat and measuring the level of privacy of a model is one of the major cha llenges in machine learning today. Identifying training data based on a trained model is a standard way of measuring the privacy risks induced by the model. We develop a novel approach to address the problem of membership inference in patte rn recognition models, relying on information provided by adversarial examples. The strategy we propose consists of measuring the magnitude of a perturbation ne cessary to build an adversarial example. Indeed, we argue that this quantity ref lects the likelihood of belonging to the training data. Extensive numerical expe riments on multivariate data and an array of state-of-the-art target models show that our method performs comparable or even outperforms state-of-the-art strate gies, but without requiring any additional training samples.
****
Raw High-Definition Radar for Multi-Task Learning
Julien Rebut, Arthur Ouaknine, Waqas Malik, Patrick Pérez; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1 7021-17030
With their robustness to adverse weather conditions and ability to measure speed s, radar sensors have been part of the automotive landscape for more than two de cades. Recent progress toward High Definition (HD) Imaging radar has driven the angular resolution below the degree, thus approaching laser scanning performance . However, the amount of data a HD radar delivers and the computational cost to estimate the angular positions remain a challenge. In this paper, we propose a n ovel HD radar sensing model, FFT-RadNet, that eliminates the overhead of computi ng the range-azimuth-Doppler 3D tensor, learning instead to recover angles from a range-Doppler spectrum. FFTRadNet is trained both to detect vehicles and to se gment free driving space. On both tasks, it competes with the most recent radar-

based models while requiring less compute and memory. Also, we collected and ann
otated 2-hour worth of raw data from synchronized automotive-grade sensors (came
ra, laser, HD radar) in various environments (city street, highway, countryside
road). This unique dataset, nick-named RADIal for "Radar, LiDAR et al.", is avai
lable at https://github.com/valeoai/RADIal.
****

Point-NeRF: Point-Based Neural Radiance Fields
Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, U
lrich Neumann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2022, pp. 5438-5448
Volumetric neural rendering methods like NeRF generate high-quality view synthes
is results but are optimized per-scene leading to prohibitive reconstruction tim
e. On the other hand, deep multi-view stereo methods can quickly reconstruct sce
ne geometry via direct network inference. Point-NeRF combines the advantages of
these two approaches by using neural 3D point clouds, with associated neural fea
tures, to model a radiance field. Point-NeRF can be rendered efficiently by aggr
egating neural point features near scene surfaces, in a ray marching-based rende
ring pipeline. Moreover, Point-NeRF can be initialized via direct inference of a
 pre-trained deep network to produce a neural point cloud; this point cloud can
be fine-tuned to surpass the visual quality of NeRF with 30X faster training tim
e. Point-NeRF can be combined with other 3D reconstruction methods and handles t
he errors and outliers in such methods via a novel pruning and growing mechanism
.
****

Contextual Debiasing for Visual Recognition With Causal Mechanisms
Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, Tao Yang; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp.
 12755-12765
As a common problem in the visual world, contextual bias means the recognition m
ay depend on the co-occurrence context rather than the objects themselves, which
 is even more severe in multi-label tasks due to multiple targets and the absenc
e of location. Although some studies have focused on tackling the problem, remov
ing the negative effect of context is still challenging because it is difficult
to obtain the representation of contextual bias. In this paper, we propose a sim
ple but effective framework employing causal inference to mitigate contextual bi
as. We first present a Structural Causal Model (SCM) clarifying the causal relat
ion among object representations, context, and predictions. Then, we develop a n
ovel Causal Context Debiasing (CCD) Module to pursue the direct effect of an ins
tance. Specifically, we adopt causal intervention to eliminate the effect of con
founder and counterfactual reasoning to obtain a Total Direct Effect (TDE) free
from the contextual bias. Note that our CCD framework is orthogonal to existing
statistical models and thus can be migrated to any other backbones. Extensive ex
periments on several multi-label classification datasets demonstrate the superio
rity of our model over other state-of-the-art baselines.
****

Complex Video Action Reasoning via Learnable Markov Logic Network
Yang Jin, Linchao Zhu, Yadong Mu; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2022, pp. 3242-3251
Profiting from the advance of deep convolutional networks, current state-of-the-
art video action recognition models have achieved remarkable progress. Neverthel
ess, most of existing models suffer from low interpretability of the predicted a
ctions. Inspired by the observation that temporally-configured human-object inte
ractions often serve as a key indicator of many actions, this work crafts an act
ion reasoning framework that performs Markov Logic Network (MLN) based probabili
stic logical inference. Crucially, we propose to encode an action by first-order
 logical rules that correspond to the temporal changes of visual relationships i
n videos. The main contributions of this work are two-fold: 1) Different from ex
isting black-box models, the proposed model simultaneously implements the locali
zation of temporal boundaries and the recognition of action categories by ground
ing the logical rules of MLN in videos. The weight associated with each such rul

e further provides an estimate of confidence. These collectively make our model more explainable and robust. 2) Instead of using hand-crafted logical rules in conventional MLN, we develop a data-driven instantiation of the MLN. In specific, a hybrid learning scheme is proposed. It combines MLN's weight learning and reinforcement learning, using the former's results as a self-critic for guiding the latter's training. Additionally, by treating actions as logical predicates, the proposed framework can also be integrated with deep models for further performance boost. Comprehensive experiments on two complex video action datasets (Charades & CAD-120) clearly demonstrate the effectiveness and explainability of our proposed method.

****

Per-Clip Video Object Segmentation

Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, Joon-Young Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1352-1361

Recently, memory-based approaches show promising results on semi-supervised video object segmentation. These methods predict object masks frame-by-frame with the help of frequently updated memory of the previous mask. Different from this per-frame inference, we investigate an alternative perspective by treating video object segmentation as clip-wise mask propagation. In this per-clip inference scheme, we update the memory with an interval and simultaneously process a set of consecutive frames (i.e. clip) between the memory updates. The scheme provides two potential benefits: accuracy gain by clip-level optimization and efficiency gain by parallel computation of multiple frames. To this end, we propose a new method tailored for the per-clip inference. Specifically, we first introduce a clip-wise operation to refine the features based on intra-clip correlation. In addition, we employ a progressive matching mechanism for efficient information-passing within a clip. With the synergy of two modules and a newly proposed per-clip based training, our network achieves state-of-the-art performance on Youtube-VOS 2018/2019 val (84.6% and 84.6%) and DAVIS 2016/2017 val (91.9% and 86.1%). Furthermore, our model shows a great speed-accuracy trade-off with varying memory update intervals, which leads to huge flexibility.

****

Exploring Set Similarity for Dense Self-Supervised Representation Learning

Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, Tongliang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16590-16599

By considering the spatial correspondence, dense self-supervised representation learning has achieved superior performance on various dense prediction tasks. However, the pixel-level correspondence tends to be noisy because of many similar misleading pixels, e.g., backgrounds. To address this issue, in this paper, we propose to explore set similarity (SetSim) for dense self-supervised representation learning. We generalize pixel-wise similarity learning to set-wise one to improve the robustness because sets contain more semantic and structure information. Specifically, by resorting to attentional features of views, we establish the corresponding set, thus filtering out noisy backgrounds that may cause incorrect correspondences. Meanwhile, these attentional features can keep the coherence of the same image across different views to alleviate semantic inconsistency. We further search the cross-view nearest neighbours of sets and employ the structured neighbourhood information to enhance the robustness. Empirical evaluations demonstrate that SetSim surpasses or is on par with state-of-the-art methods on object detection, keypoint detection, instance segmentation, and semantic segmentation.

****

Coarse-To-Fine Feature Mining for Video Semantic Segmentation

Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3126-3137

The contextual information plays a core role in semantic segmentation. As for video semantic segmentation, the contexts include static contexts and motional con

texts, corresponding to static content and moving content in a video clip, respe
ctively. The static contexts are well exploited in image semantic segmentation b
y learning multi-scale and global/long-range features. The motional contexts are
 studied in previous video semantic segmentation. However, there is no research
about how to simultaneously learn static and motional contexts which are highly
correlated and complementary to each other. To address this problem, we propose
a Coarse-to-Fine Feature Mining (CFFM) technique to learn a unified presentation
 of static contexts and motional contexts. This technique consists of two parts:
 coarse-to-fine feature assembling and cross-frame feature mining. The former op
eration prepares data for further processing, enabling the subsequent joint lear
ning of static and motional contexts. The latter operation mines useful informat
ion/contexts from the sequential frames to enhance the video contexts of the fea
tures of the target frame. The enhanced features can be directly applied for the
 final prediction. Experimental results on popular benchmarks demonstrate that t
he proposed CFFM performs favorably against state-of-the-art methods for video s
emantic segmentation. Our implementation is available at https://github.com/Guol
eiSun/VSS-CFFM
****

ONCE-3DLanes: Building Monocular 3D Lane Detection
Fan Yan, Ming Nie, Xinyue Cai, Jianhua Han, Hang Xu, Zhen Yang, Chaoqiang Ye, Ya
nwei Fu, Michael Bi Mi, Li Zhang; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2022, pp. 17143-17152
We present ONCE-3DLanes, a real-world autonomous driving dataset with lane layou
t annotation in 3D space. Conventional 2D lane detection from a monocular image
yields poor performance of following planning and control tasks in autonomous dr
iving due to the case of uneven road. Predicting the 3D lane layout is thus nece
ssary and enables effective and safe driving. However, existing 3D lane detectio
n datasets are either unpublished or synthesized from a simulated environment, s
everely hampering the development of this field. In this paper, we take steps to
wards addressing these issues. By exploiting the explicit relationship between p
oint clouds and image pixels, a dataset annotation pipeline is designed to autom
atically generate high-quality 3D lane locations from 2D lane annotations in 211
K road scenes. In addition, we present an extrinsic-free, anchor-free method, ca
lled SALAD, regressing the 3D coordinates of lanes in image view without convert
ing the feature map into the bird's-eye view (BEV). To facilitate future researc
h on 3D lane detection, we benchmark the dataset and provide a novel evaluation
metric, performing extensive experiments of both existing approaches and our pro
posed method. The aim of our work is to revive the interest of 3D lane detection
 in a real-world scenario. We believe our work can lead to expected and unexpect
ed innovations in both academia and industry.
****

Weakly but Deeply Supervised Occlusion-Reasoned Parametric Road Layouts
Buyu Liu, Bingbing Zhuang, Manmohan Chandraker; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17000-17009
We propose an end-to-end network that takes a single perspective RGB image of a
complex road scene as input, to produce occlusion-reasoned layouts in perspectiv
e space as well as a parametric bird's-eye-view (BEV) space. In contrast to prio
r works that require dense supervision such as semantic labels in perspective vi
ew, our method only requires human annotations for parametric attributes that ar
e cheaper and less ambiguous to obtain. To solve this challenging task, our desi
gn is comprised of modules that incorporate inductive biases to learn occlusion-
reasoning, geometric transformation and semantic abstraction, where each module
may be supervised by appropriately transforming the parametric annotations. We d
emonstrate how our design choices and proposed deep supervision help achieve mea
ningful representations and accurate predictions. We validate our approach on tw
o public datasets, KITTI and NuScenes, to achieve state-of-the-art results with
considerably less human supervision.
****

Compressing Models With Few Samples: Mimicking Then Replacing
Huanyu Wang, Junjie Liu, Xin Ma, Yang Yong, Zhenhua Chai, Jianxin Wu; Proceeding

s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 701-710
Few-sample compression aims to compress a big redundant model into a small compact one with only few samples. If we fine-tune models with these limited few samples directly, models will be vulnerable to overfit and learn almost nothing. Hence, previous methods optimize the compressed model layer-by-layer and try to make every layer have the same outputs as the corresponding layer in the teacher model, which is cumbersome. In this paper, we propose a new framework named Mimicking then Replacing (MiR) for few-sample compression, which firstly urges the pruned model to output the same features as the teacher's in the penultimate layer, and then replaces teacher's layers before penultimate with a well-tuned compact one. Unlike previous layer-wise reconstruction methods, our MiR optimizes the entire network holistically, which is not only simple and effective, but also unsupervised and general. MiR outperforms previous methods with large margins. Codes is available at https://github.com/cjnjuwhy/MiR.
****

FedCor: Correlation-Based Active Client Selection Strategy for Heterogeneous Federated Learning
Minxue Tang, Xuefei Ning, Yitu Wang, Jingwei Sun, Yu Wang, Hai Li, Yiran Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10102-10111
Client-wise data heterogeneity is one of the major issues that hinder effective training in federated learning (FL). Since the data distribution on each client may vary dramatically, the client selection strategy can significantly influence the convergence rate of the FL process. Active client selection strategies are popularly proposed in recent studies. However, they neglect the loss correlations between the clients and achieve only marginal improvement compared to the uniform selection strategy. In this work, we propose FedCor---an FL framework built on a correlation-based client selection strategy, to boost the convergence rate of FL. Specifically, we first model the loss correlations between the clients with a Gaussian Process (GP). Based on the GP model, we derive a client selection strategy with a significant reduction of expected global loss in each round. Besides, we develop an efficient GP training method with a low communication overhead in the FL scenario by utilizing the covariance stationarity. Our experimental results show that compared to the state-of-the-art method, FedCorr can improve the convergence rates by 34%   99% and 26%   51% on FMNIST and CIFAR-10, respectively.
****

Modulated Contrast for Versatile Image Synthesis
Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18280-18290
Perceiving the similarity between images has been a long-standing and fundamental problem underlying various visual generation tasks. Predominant approaches measure the inter-image distance by computing pointwise absolute deviations, which tends to estimate the median of instance distributions and leads to blurs and artifacts in the generated images. This paper presents MoNCE, a versatile metric that introduces image contrast to learn a calibrated metric for the perception of multifaceted inter-image distances. Unlike vanilla contrast which indiscriminately pushes negative samples from the anchor regardless of their similarity, we propose to re-weight the pushing force of negative samples adaptively according to their similarity to the anchor, which facilitates the contrastive learning from informative negative samples. Since multiple patch-level contrastive objectives are involved in image distance measurement, we introduce optimal transport in MoNCE to modulate the pushing force of negative samples collaboratively across multiple contrastive objectives. Extensive experiments over multiple image translation tasks show that the proposed MoNCE outperforms various prevailing metrics substantially.
****

PokeBNN: A Binary Pursuit of Lightweight Accuracy

Yichi Zhang, Zhiru Zhang, Lukasz Lew; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12475-12485
Optimization of Top-1 ImageNet promotes enormous networks that may be impractical in inference settings. Binary neural networks (BNNs) have the potential to significantly lower the compute intensity but existing models suffer from low quality. To overcome this deficiency, we propose PokeConv, a binary convolution block which improves quality of BNNs by techniques such as adding multiple residual paths, and tuning the activation function. We apply it to ResNet-50 and optimize ResNet's initial convolutional layer which is hard to binarize. We name the resulting network family PokeBNN. These techniques are chosen to yield favorable improvements in both top-1 accuracy and the network's cost. In order to enable joint optimization of the cost together with accuracy, we define arithmetic computation effort (ACE), a hardware- and energy-inspired cost metric for quantized and binarized networks. We also identify a need to optimize an under-explored hyper-parameter controlling the binarization gradient approximation. We establish a new, strong state-of-the-art (SOTA) on top-1 accuracy together with commonly-used CPU64 cost, ACE cost and network size metrics. ReActNet-Adam, the previous SOTA in BNNs, achieved a 70.5% top-1 accuracy with 7.9 ACE. A small variant of PokeBNN achieves 70.5% top-1 with 2.6 ACE, more than 3x reduction in cost; a larger PokeBNN achieves 75.6% top-1 with 7.8 ACE, more than 5% improvement in accuracy without increasing the cost. PokeBNN implementation in JAX / Flax and reproduction instructions are open sourced. Source code and reproduction instructions are available in AQT repository: https://github.com/google/aqt.
****

HumanNeRF: Efficiently Generated Human Radiance Field From Sparse Inputs
Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7743-7753
Recent neural human representations can produce high-quality multi-view rendering but require using dense multi-view inputs and costly training. They are hence largely limited to static models as training each frame is infeasible. We present HumanNeRF - a neural representation with efficient generalization ability - for high-fidelity free-view synthesis of dynamic humans. Analogous to how IBRNet assists NeRF by avoiding per-scene training, HumanNeRF employs an aggregated pixel-alignment feature across multi-view inputs along with a pose embedded non-rigid deformation field for tackling dynamic motions. The raw HumanNeRF can already produce reasonable rendering on sparse video inputs of unseen subjects and camera settings. To further improve the rendering quality, we augment our solution with in-hour scene-specific fine-tuning, and an appearance blending module for combining the benefits of both neural volumetric rendering and neural texture blending. Extensive experiments on various multi-view dynamic human datasets demonstrate effectiveness of our approach in synthesizing photo-realistic free-view humans under challenging motions and with very sparse camera view inputs.
****

Zoom in and Out: A Mixed-Scale Triplet Network for Camouflaged Object Detection
Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2160-2170
The recently proposed camouflaged object detection (COD) attempts to segment objects that are visually blended into their surroundings, which is extremely complex and difficult in real-world scenarios. Apart from high intrinsic similarity between the camouflaged objects and their background, the objects are usually diverse in scale, fuzzy in appearance, and even severely occluded. To deal with these problems, we propose a mixed-scale triplet network, ZoomNet, which mimics the behavior of humans when observing vague images, i.e., zooming in and out. Specifically, our ZoomNet employs the zoom strategy to learn the discriminative mixed-scale semantics by the designed scale integration unit and hierarchical mixed-scale unit, which fully explores imperceptible clues between the candidate objects and background surroundings. Moreover, considering the uncertainty and ambiguity derived from indistinguishable textures, we construct a simple yet effective

regularization constraint, uncertainty-aware loss, to promote the model to accurately produce predictions with higher confidence in candidate regions. Without bells and whistles, our proposed highly task-friendly model consistently surpasses the existing 23 state-of-the-art methods on four public datasets. Besides, the superior performance over the recent cutting-edge models on the SOD task also verifies the effectiveness and generality of our model. The code will be available at https://github.com/lartpang/ZoomNet.

****

Identifying Ambiguous Similarity Conditions via Semantic Matching

Han-Jia Ye, Yi Shi, De-Chuan Zhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16610-16619

Rich semantics inside an image result in its ambiguous relationship with others, i.e., two images could be similar in one condition but dissimilar in another. Given triplets like "aircraft" is similar to "bird" than "train", Weakly Supervised Conditional Similarity Learning (WS-CSL) learns multiple embeddings to match semantic conditions without explicit condition labels such as "can fly". However, similarity relationships in a triplet are uncertain except providing a condition. For example, the previous comparison becomes invalid once the conditional label changes to "is vehicle". To this end, we introduce a novel evaluation criterion by predicting the comparison's correctness after assigning the learned embeddings to their optimal conditions, which measures how much WS-CSL could cover latent semantics as the supervised model. Furthermore, we propose the Distance Induced Semantic COndition VERification Network (DiscoverNET), which characterizes the instance-instance and triplets-condition relations in a "decompose-and-fuse" manner. To make the learned embeddings cover all semantics, DiscoverNET utilizes a set module or an additional regularizer over the correspondence between a triplet and a condition. DiscoverNET achieves state-of-the-art performance on benchmarks like UT-Zappos-50k and Celeb-A w.r.t. different criteria.

****

MISF: Multi-Level Interactive Siamese Filtering for High-Fidelity Image Inpainting

Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1869-1878

Although achieving significant progress, existing deep generative inpainting methods still show low generalization across different scenes. As a result, the generated images usually contain artifacts or the filled pixels differ greatly from the ground truth, making them far from real-world applications. Image-level predictive filtering is a widely used restoration technique by predicting suitable kernels adaptively according to different input scenes. Inspired by this inherent advantage, we explore the possibility of addressing image inpainting as a filtering task. To this end, we first study the advantages and challenges of the image-level predictive filtering for inpainting: the method can preserve local structures and avoid artifacts but fails to fill large missing areas. Then, we propose the semantic filtering by conducting filtering on deep feature level, which fills the missing semantic information but fails to recover the details. To address the issues while adopting the respective advantages, we propose a novel filtering technique, i.e., Multi-level Interactive Siamese Filtering (MISF) containing two branches: kernel prediction branch (KPB) and semantic & image filtering branch (SIFB). These two branches are interactively linked: SIFB provides multi-level features for KPB while KPB predicts dynamic kernels for SIFB. As a result, the final method takes the advantage of effective semantic & image-level filling for high-fidelity inpainting. Moreover, we discuss the relationship between MISF and the naive encoder-decoder-based inpainting, inferring that MISF provides novel dynamic convolutional operations to enhance the high generalization capability across scenes. We validate our method on three challenging datasets, i.e., Dunhuang, Places2, and CelebA. Our method outperforms state-of-the-art baselines on four metrics, i.e., L1, PSNR, SSIM, and LPIPS.

****

Cascade Transformers for End-to-End Person Search

Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, Brian Clipp; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7267-7276

The goal of person search is to localize a target person from a gallery set of scene images, which is extremely challenging due to large scale variations, pose/viewpoint changes, and occlusions. In this paper, we propose the Cascade Occluded Attention Transformer (COAT) for end-to-end person search. Our three-stage cascade design focuses on detecting people in the first stage, while later stages simultaneously and progressively refine the representation for person detection and re-identification. At each stage the occluded attention transformer applies tighter intersection over union thresholds, forcing the network to learn coarse-to-fine pose/scale invariant features. Meanwhile, we calculate each detection's occluded attention to differentiate a person's tokens from other people or the background. In this way, we simulate the effect of other objects occluding a person of interest at the token-level. Through comprehensive experiments, we demonstrate the benefits of our method by achieving state-of-the-art performance on two benchmark datasets.

****

MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection

Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, Eun-Sol Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19578-19587

Human-Object Interaction (HOI) detection is the task of identifying a set of <human, object, interaction> triplets from an image. Recent work proposed transformer encoder-decoder architectures that successfully eliminated the need for many hand-designed components in HOI detection through end-to-end training. However, they are limited to single-scale feature resolution, providing suboptimal performance in scenes containing humans, objects, and their interactions with vastly different scales and distances. To tackle this problem, we propose a Multi-Scale TRansformer (MSTR) for HOI detection powered by two novel HOI-aware deformable attention modules called Dual-Entity attention and Entity-conditioned Context attention. While existing deformable attention comes at a huge cost in HOI detection performance, our proposed attention modules of MSTR learn to effectively attend to sampling points that are essential to identify interactions. In experiments, we achieve the new state-of-the-art performance on two HOI detection benchmarks.

****

LSVC: A Learning-Based Stereo Video Compression Framework

Zhenghao Chen, Guo Lu, Zhihao Hu, Shan Liu, Wei Jiang, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6073-6082

In this work, we propose the first end-to-end optimized framework for compressing automotive stereo videos (i.e., stereo videos from autonomous driving applications) from both left and right views. Specifically, when compressing the current frame from each view, our framework reduces temporal redundancy by performing motion compensation using the reconstructed intra-view adjacent frame and at the same time exploits binocular redundancy by conducting disparity compensation using the latest reconstructed cross-view frame. Moreover, to effectively compress the introduced motion and disparity offsets for better compensation, we further propose two novel schemes called motion residual compression and disparity residual compression to respectively generate the predicted motion offset and disparity offset from the previously compressed motion offset and disparity offset, such that we can more effectively compress residual offset information for better bit-rate saving. Overall, the entire framework is implemented by the fully-differentiable modules and can be optimized in an end-to-end manner. Our comprehensive experiments on three automotive stereo video benchmarks Cityscapes, KITTI 2012 and KITTI 2015 demonstrate that our proposed framework outperforms the learning-based single-view video codec and the traditional hand-crafted multi-view video codec.

****

How Do You Do It? Fine-Grained Action Understanding With Pseudo-Adverbs
Hazel Doughty, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2022, pp. 13832-13842
We aim to understand how actions are performed and identify subtle differences, such as 'fold firmly' vs. 'fold gently'. To this end, we propose a method which recognizes adverbs across different actions. However, such fine-grained annotati ons are difficult to obtain and their long-tailed nature makes it challenging to recognize adverbs in rare action-adverb compositions. Our approach therefore us es semi-supervised learning with multiple adverb pseudo-labels to leverage video s with only action labels. Combined with adaptive thresholding of these pseudo-a dverbs we are able to make efficient use of the available data while tackling th e long-tailed distribution. Additionally, we gather adverb annotations for three existing video retrieval datasets, which allows us to introduce the new tasks o f recognizing adverbs in unseen action-adverb compositions and unseen domains. E xperiments demonstrate the effectiveness of our method,which outperforms prior w ork in recognizing adverbs and semi-supervised works adapted for adverb recognit ion. We also show how adverbs can relate fine-grained actions.
****

InsetGAN for Full-Body Image Generation
Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J. Mitra, Peter Wonka, Jingwan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2022, pp. 7723-7732
While GANs can produce photo-realistic images in ideal conditions for certain do mains, the generation of full-body human images remains difficult due to the div ersity of identities, hairstyles, clothing, and the variance in pose. Instead of modeling this complex domain with a single GAN, we propose a novel method to co mbine multiple pretrained GANs, where one GAN generates a global canvas (e.g., h uman body) and a set of specialized GANs, or insets, focus on different parts (e .g., faces, shoes) that can be seamlessly inserted onto the global canvas. We mo del the problem as jointly exploring the respective latent spaces such that the generated images can be combined, by inserting the parts from the specialized ge nerators onto the global canvas, without introducing seams. We demonstrate the s etup by combining a full body GAN with a dedicated high-quality face GAN to prod uce plausible-looking humans. We evaluate our results with quantitative metrics and user studies.
****

DetectorDetective: Investigating the Effects of Adversarial Examples on Object D etectors
Sivapriya Vellaichamy, Matthew Hull, Zijie J. Wang, Nilaksh Das, ShengYun Peng, Haekyu Park, Duen Horng (Polo) Chau; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2022, pp. 21484-21491
With deep learning based systems performing exceedingly well in many vision-rela ted tasks, a major concern with their widespread deployment especially in safety -critical applications is their susceptibility to adversarial attacks. We propos e DetectorDetective, an interactive visual tool that aims to help users better u nderstand the behaviors of a model as adversarial images journey through an obje ct detector. DetectorDetective enables users to easily learn about how the three key modules of the Faster R-CNN object detector -- Feature Pyramidal Network, R egion Proposal Network, and Region Of Interest Head--respond to a user-selected benign image and its adversarial version. Visualizations about the progressive c hanges in the intermediate features among such modules help users gain insights into the impact of adversarial attacks, and perform side-by-side comparisons bet ween the benign and adversarial responses. Furthermore, DetectorDetective displa ys saliency maps for the input images to comparatively highlight image regions t hat contribute to attack success. DetectorDetective complements adversarial mach ine learning research on object detection by providing a user-friendly interacti ve tool for inspecting and understanding model responses. DetectorDetective is a vailable at the following public demo link: https://poloclub.github.io/detector-detective. A video demo is available at https://youtu.be/5C3Klh87CZI.
****

SOMSI: Spherical Novel View Synthesis With Soft Occlusion Multi-Sphere Images
Tewodros Habtegebrial, Christiano Gava, Marcel Rogge, Didier Stricker, Varun Jam pani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2022, pp. 15725-15734
Spherical novel view synthesis (SNVS) is the task of estimating 360 views at dyn amic novel views given a set of 360 input views. Prior arts learn multi-sphere i mage (MSI) representations that enables fast rendering times but are only limite d to modelling low-dimensional color values. Modelling high-dimensional appearan ce features in MSI can result in better view synthesis, but it is not feasible t o represent high-dimensional features in a large number (>64) of MSI spheres. We propose a novel MSI representation called Soft Occlusion MSI (SOMSI) that enabl es modelling high-dimensional appearance features in MSI while retaining the fas t rendering times of a standard MSI. Our key insight is to model appearance feat ures in a smaller set (e.g. 3) of occlusion levels instead of larger number of M SI levels. Experiments on both synthetic and real-world scenes demonstrate that using SOMSI can provide a good balance between accuracy and runtime. SOMSI can p roduce considerably better results compared to MSI based MODS, while having simi lar fast rendering time. SOMSI view synthesis quality is on-par with state-of-th e-art NeRF like model while being 2 orders of magnitude faster. For code, additi onal results and data, please visit https://tedyhabtegebrial.github.io/somsi.
****

EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embeddi ng Matching
Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, Zheng-Jun Zha ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2022, pp. 17929-17938
Current metrics for video captioning are mostly based on the text-level comparis on between reference and candidate captions. However, they have some insuperable drawbacks, e.g., they cannot handle videos without references, and they may res ult in biased evaluation due to the one-to-many nature of video-to-text and the neglect of visual relevance. From the human evaluator's viewpoint, a high-qualit y caption should be consistent with the provided video, but not necessarily be s imilar to the reference in literal or semantics. Inspired by human evaluation, w e propose EMScore (Embedding Matching-based score), a novel reference-free metri c for video captioning, which directly measures similarity between video and can didate captions. Benefiting from the recent development of large-scale pre-train ing models, we exploit a well pre-trained vision-language model to extract visua l and linguistic embeddings for computing EMScore. Specifically, EMScore combine s matching scores of both coarse-grained (video and caption) and fine-grained (f rames and words) levels, which takes the overall understanding and detailed char acteristics of the video into account. Furthermore, considering the potential in formation gain, EMScore can be flexibly extended to the conditions where human-l abeled references are available. Last but not least, we collect VATEX-EVAL and A ctivityNet-FOIl datasets to systematically evaluate the existing metrics. VATEX-EVAL experiments demonstrate that EMScore has higher human correlation and lower reference dependency. ActivityNet-FOIL experiment verifies that EMScore can eff ectively identify hallucinating captions. Code and datasets are available at htt ps://github.com/shiyaya/emscore.
****

SNR-Aware Low-Light Image Enhancement
Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17714-177 24
This paper presents a new solution for low-light image enhancement by collective ly exploiting Signal-to-Noise-Ratio-aware transformers and convolutional models to dynamically enhance pixels with spatial-varying operations. They are long-ran ge operations for image regions of extremely low Signal-to-Noise-Ratio (SNR) and short-range operations for other regions. We propose to take an SNR prior to gu ide the feature fusion and formulate the SNR-aware transformer with a new self-a ttention model to avoid tokens from noisy image regions of very low SNR. Extensi

ve experiments show that our framework consistently achieves better performance than SOTA approaches on seven representative benchmarks with the same structure. Also, we conducted a large-scale user study with 100 participants to verify the superior perceptual quality of our results.
****

3D Common Corruptions and Data Augmentation
O■uzhan Fatih Kar, Teresa Yeo, Andrei Atanov, Amir Zamir; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18963-18974
We introduce a set of image transformations that can be used as corruptions to evaluate the robustness of models as well as data augmentation mechanisms for training neural networks. The primary distinction of the proposed transformations is that, unlike existing approaches such as Common Corruptions, the geometry of the scene is incorporated in the transformations -- thus leading to corruptions that are more likely to occur in the real world. We also introduce a set of semantic corruptions (e.g. natural object occlusions). We show these transformations are 'efficient' (can be computed on-the-fly), 'extendable' (can be applied on most image datasets), expose vulnerability of existing models, and can effectively make models more robust when employed as '3D data augmentation' mechanisms. The evaluations on several tasks and datasets suggest incorporating 3D information into benchmarking and training opens up a promising direction for robustness research.
****

PoseTriplet: Co-Evolving 3D Human Pose Estimation, Imitation, and Hallucination Under Self-Supervision
Kehong Gong, Bingbing Li, Jianfeng Zhang, Tao Wang, Jing Huang, Michael Bi Mi, Jiashi Feng, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11017-11027
Existing self-supervised 3D human pose estimation schemes have largely relied on weak supervisions like consistency loss to guide the learning, which, inevitably, leads to inferior results in real-world scenarios with unseen poses. In this paper, we propose a novel self-supervised approach that allows us to explicitly generate 2D-3D pose pairs for augmenting supervision, through a self-enhancing dual-loop learning framework. This is made possible via introducing a reinforcement-learning-based imitator, which is learned jointly with a pose estimator alongside a pose hallucinator; the three components form two loops during the training process, complementing and strengthening one another. Specifically, the pose estimator transforms an input 2D pose sequence to a low-fidelity 3D output, which is then enhanced by the imitator that enforces physical constraints. The refined 3D poses are subsequently fed to the hallucinator for producing even more diverse data, which are, in turn, strengthened by the imitator and further utilized to train the pose estimator. Such a co-evolution scheme, in practice, enables training a pose estimator on self-generated motion data without relying on any given 3D data. Extensive experiments across various benchmarks demonstrate that our approach yields encouraging results significantly outperforming the state of the art and, in some cases, even on par with results of fully-supervised methods. Notably, it achieves 89.1% 3D PCK on MPI-INF-3DHP under self-supervised cross-dataset evaluation setup, improving upon the previous best self-supervised method by 8.6%. Code is available at https://github.com/Garfield-kh/PoseTriplet.
****

Injecting Semantic Concepts Into End-to-End Image Captioning
Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18009-18019
Tremendous progress has been made in recent years in developing better image captioning models, yet most of them rely on a separate object detector to extract regional features. Recent vision-language studies are shifting towards the detector-free trend by leveraging grid representations for more flexible model training and faster inference speed. However, such development is primarily focused on image understanding tasks, and remains less investigated for the caption generat

ion task. In this paper, we are concerned with a better-performing detector-free image captioning model, and propose a pure vision transformer-based image captioning model, dubbed as ViTCAP, in which grid representations are used without extracting the regional features. For improved performance, we introduce a novel Concept Token Network (CTN) to predict the semantic concepts and then incorporate them into the end-to-end captioning. In particular, the CTN is built on the basis of a vision transformer and is designed to predict the concept tokens through a classification task, from which the rich semantic information contained greatly benefits the captioning task. Compared with the previous detector-based models, ViTCAP drastically simplifies the architectures and at the same time achieves competitive performance on various challenging image captioning datasets. In particular, ViTCAP reaches 138.1 CIDEr scores on COCO-caption Karpathy-split, 93.8 and 108.6 CIDEr scores on nocaps and Google-CC captioning datasets, respectively.

****

An Efficient Training Approach for Very Large Scale Face Recognition

Kai Wang, Shuo Wang, Panpan Zhang, Zhipeng Zhou, Zheng Zhu, Xiaobo Wang, Xiaojiang Peng, Baigui Sun, Hao Li, Yang You; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4083-4092

Face recognition has achieved significant progress in deep learning era due to the ultra-large-scale and welllabeled datasets. However, training on the outsize datasets is time-consuming and takes up a lot of hardware resource. Therefore, designing an efficient training approach is indispensable. The heavy computational and memory costs mainly result from the million-level dimensionality of thefully connected (FC) layer. To this end, we propose a novel training approach, termed Faster Face Classification (F2C), to alleviate time and cost without sacrificing the performance. This method adopts Dynamic Class Pool (DCP) for storing and updating the identities' features dynamically, which could be regarded as a substitute for the FC layer. DCP is efficiently time-saving and cost-saving, as its smaller size with the independence from the whole face identities together. We further validate the proposed F2C method across several face benchmarks and private datasets, and display comparable results, meanwhile the speed is faster than state-of-the-art FC-based methods in terms of recognition accuracy and hardware costs. Moreover, our method is further improved by a well-designed dual data loader including indentity-based and instancebased loaders, which makes it more efficient for the updating DCP parameters.

****

Long-Term Video Frame Interpolation via Feature Propagation

Dawit Mureja Argaw, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3543-3552

Video frame interpolation (VFI) works generally predict intermediate frame(s) by first estimating the motion between inputs and then warping the inputs to the target time with the estimated motion. This approach, however, is not optimal when the temporal distance between the input sequence increases as existing motion estimation modules cannot effectively handle large motions. Hence, VFI works perform well for small frame gaps and perform poorly as the frame gap increases. In this work, we propose a novel framework to address this problem. We argue that when there is a large gap between inputs, instead of estimating imprecise motion that will eventually lead to inaccurate interpolation, we can safely propagate from one side of the input up to a reliable time frame using the other input as a reference. Then, the rest of the intermediate frames can be interpolated using standard approaches as the temporal gap is now narrowed. To this end, we propose a propagation network (PNet) by extending the classic feature-level forecasting with a novel motion-to-feature approach. To be thorough, we adopt a simple interpolation model along with PNet as our full model and design a simple procedure to train the full model in an end-to-end manner. Experimental results on several benchmark datasets confirm the effectiveness of our method for long-term VFI compared to state-of-the-art approaches.

****

Coarse-To-Fine Q-Attention: Efficient Learning for Visual Robotic Manipulation v

ia Discretisation
Stephen James, Kentaro Wada, Tristan Laidlow, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13739-13748
We present a coarse-to-fine discretisation method that enables the use of discrete reinforcement learning approaches in place of unstable and data-inefficient actor-critic methods in continuous robotics domains. This approach builds on the recently released ARM algorithm, which replaces the continuous next-best pose agent with a discrete one, with coarse-to-fine Q-attention. Given a voxelised scene, coarse-to-fine Q-attention learns what part of the scene to 'zoom' into. When this 'zooming' behaviour is applied iteratively, it results in a near-lossless discretisation of the translation space, and allows the use of a discrete action, deep Q-learning method. We show that our new coarse-to-fine algorithm achieves state-of-the-art performance on several difficult sparsely rewarded RLBench vision-based robotics tasks, and can train real-world policies, tabula rasa, in a matter of minutes, with as little as 3 demonstrations.
****

Event-Aided Direct Sparse Odometry
Javier Hidalgo-Carrió, Guillermo Gallego, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5781-5790
We introduce EDS, a direct monocular visual odometry using events and frames. Our algorithm leverages the event generation model to track the camera motion in the blind time between frames. The method formulates a direct probabilistic approach of observed brightness increments. Per-pixel brightness increments are predicted using a sparse number of selected 3D points and are compared to the events via the brightness increment error to estimate camera motion. The method recovers a semi-dense 3D map using photometric bundle adjustment. EDS is the first method to perform 6-DOF VO using events and frames with a direct approach. By design it overcomes the problem of changing appearance in indirect methods. Our results outperform all previous event-based odometry solutions. We also show that, for a target error performance, EDS can work at lower frame rates than state-of-the-art frame-based VO solutions. This opens the door to low-power motion-tracking applications where frames are sparingly triggered "on demand" and our method tracks the motion in between. We release code and datasets to the public.
****

Group Contextualization for Video Recognition
Yanbin Hao, Hao Zhang, Chong-Wah Ngo, Xiangnan He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 928-938
Learning discriminative representation from the complex spatio-temporal dynamic space is essential for video recognition. On top of those stylized spatio-temporal computational units, further refining the learnt feature with axial contexts is demonstrated to be promising in achieving this goal. However, previous works generally focus on utilizing a single kind of contexts to calibrate entire feature channels and could hardly apply to deal with diverse video activities. The problem can be tackled by using pair-wise spatio-temporal attentions to recompute feature response with cross-axis contexts at the expense of heavy computations. In this paper, we propose an efficient feature refinement method that decomposes the feature channels into several groups and separately refines them with different axial contexts in parallel. We refer this lightweight feature calibration as group contextualization (GC). Specifically, we design a family of efficient element-wise calibrators, i.e., ECal-G/S/T/L, where their axial contexts are information dynamics aggregated from other axes either globally or locally, to contextualize feature channel groups. The GC module can be densely plugged into each residual layer of the off-the-shelf video networks. With little computational overhead, consistent improvement is observed when plugging in GC on different networks. By utilizing calibrators to embed feature with four different kinds of contexts in parallel, the learnt representation is expected to be more resilient to diverse types of activities. On videos with rich temporal variations, empirically GC can boost the performance of 2D-CNN (e.g., TSN and TSM) to a level comparab

le to the state-of-the-art video networks. Code is available at https://github.com/haoyanbin918/Group-Contextualization.
****

Single-Domain Generalized Object Detection in Urban Scene via Cyclic-Disentangled Self-Distillation
Aming Wu, Cheng Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 847-856
In this paper, we are concerned with enhancing the generalization capability of object detectors. And we consider a realistic yet challenging scenario, namely Single-Domain Generalized Object Detection (Single-DGOD), which aims to learn an object detector that performs well on many unseen target domains with only one source domain for training. Towards Single-DGOD, it is important to extract domain-invariant representations (DIR) containing intrinsical object characteristics, which is beneficial for improving the robustness for unseen domains. Thus, we present a method, i.e., cyclic-disentangled self-distillation, to disentangle DIR from domain-specific representations without the supervision of domain-related annotations (e.g., domain labels). Concretely, a cyclic-disentangled module is first proposed to cyclically extract DIR from the input visual features. Through the cyclic operation, the disentangled ability can be promoted without the reliance on domain-related annotations. Then, taking the DIR as the teacher, we design a self-distillation module to further enhance the generalization ability. In the experiments, our method is evaluated in urban-scene object detection. Experimental results of five weather conditions show that our method obtains a significant performance gain over baseline methods. Particularly, for the night-sunny scene, our method outperforms baselines by 3%, which indicates that our method is instrumental in enhancing generalization ability. Data and code are available at https://github.com/AmingWu/Single-DGOD.
****

Visual Abductive Reasoning
Chen Liang, Wenguan Wang, Tianfei Zhou, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15565-15575
Abductive reasoning seeks the likeliest possible explanation for partial observations. Although abduction is frequently employed in human daily reasoning, it is rarely explored in computer vision literature. In this paper, we propose a new task and dataset, Visual Abductive Reasoning (VAR), for examining abductive reasoning ability of machine intelligence in everyday visual situations. Given an incomplete set of visual events, AI systems are required to not only describe what is observed, but also infer the hypothesis that can best explain the visual premise. Based on our large-scale VAR dataset, we devise a strong baseline model, Reasoner (causal-and-cascaded reasoning Transformer). First, to capture the causal structure of the observations, a contextualized directional position embedding strategy is adopted in the encoder, that yields discriminative representations for the premise and hypothesis. Then, multiple decoders are cascaded to generate and progressively refine the premise and hypothesis sentences. The prediction scores of the sentences are used to guide cross-sentence information flow in the cascaded reasoning procedure. Our VAR benchmarking results show that Reasoner surpasses many famous video-language models, while still being far behind human performance. This work is expected to foster future efforts in the reasoning-beyond-observation paradigm.
****

L2G: A Simple Local-to-Global Knowledge Transfer Framework for Weakly Supervised Semantic Segmentation
Peng-Tao Jiang, Yuqi Yang, Qibin Hou, Yunchao Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16886-16896
Mining precise class-aware attention maps, a.k.a, class activation maps, is essential for weakly supervised semantic segmentation. In this paper, we present L2G, a simple online local-to-global knowledge transfer framework for high-quality object attention mining. We observe that classification models can discover object regions with more details when replacing the input image with its local patch

es. Taking this into account, we first leverage a local classification network t
o extract attentions from multiple local patches randomly cropped from the input
 image. Then, we utilize a global network to learn complementary attention knowl
edge across multiple local attention maps online. Our framework conducts the glo
bal network to learn the captured rich object detail knowledge from a global vie
w and thereby produces high-quality attention maps that can be directly used as
pseudo annotations for semantic segmentation networks. Experiments show that our
 method attains 72.1% and 44.2% mIoU scores on the validation set of PASCAL VOC
2012 and MS COCO 2014, respectively, setting new state-of-the-art records. Code
is available at https://github.com/PengtaoJiang/L2G.
****

Rethinking Bayesian Deep Learning Methods for Semi-Supervised Volumetric Medical
 Image Segmentation
Jianfeng Wang, Thomas Lukasiewicz; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2022, pp. 182-190
Recently, several Bayesian deep learning methods have been proposed for semi-sup
ervised medical image segmentation. Although they have achieved promising result
s on medical benchmarks, some problems are still existing. Firstly, their overal
l architectures belong to the discriminative models, and hence, in the early sta
ge of training, they only use labeled data for training, which might make them o
verfit to the labeled data. Secondly, in fact, they are only partially based on
Bayesian deep learning, as their overall architectures are not designed under th
e Bayesian framework. However, unifying the overall architecture under the Bayes
ian perspective can make the architecture have a rigorous theoretical basis, so
that each part of the architecture can have a clear probabilistic interpretation
. Therefore, to solve the problems, we propose a new generative Bayesian deep le
arning (GBDL) architecture. GBDL belongs to the generative models, whose target
is to estimate the joint distribution of input medical volumes and their corresp
onding labels. Estimating the joint distribution implicitly involves the distrib
ution of data, so both labeled and unlabeled data can be utilized in the early s
tage of training, which alleviates the potential overfitting problem. Besides, G
BDL is completely designed under the Bayesian framework, and thus we give its fu
ll Bayesian formulation, which lays a theoretical probabilistic foundation for o
ur architecture. Extensive experiments show that our GBDL outperforms previous s
tate-of-the-art methods in terms of four commonly used evaluation indicators on
three public medical datasets.
****

Continual Learning With Lifelong Vision Transformer
Zhen Wang, Liu Liu, Yiqun Duan, Yajing Kong, Dacheng Tao; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17
1-181
Continual learning methods aim at training a neural network from sequential data
 with streaming labels, relieving catastrophic forgetting. However, existing met
hods are based on and designed for convolutional neural networks (CNNs), which h
ave not utilized the full potential of newly emerged powerful vision transformer
s. In this paper, we propose a novel attention-based framework Lifelong Vision T
ransformer (LVT), to achieve a better stability-plasticity trade-off for continu
al learning. Specifically, an inter-task attention mechanism is presented in LVT
, which implicitly absorbs the previous tasks' information and slows down the dr
ift of important attention between previous tasks and the current task. LVT desi
gns a dual-classifier structure that independently injects new representation to
 avoid catastrophic interference and accumulates the new and previous knowledge
in a balanced manner to improve the overall performance. Moreover, we develop a
confidence-aware memory update strategy to deepen the impression of the previous
 tasks. The extensive experimental results show that our approach achieves state
-of-the-art performance with even fewer parameters on continual learning benchma
rks.
****

MPViT: Multi-Path Vision Transformer for Dense Prediction
Youngwan Lee, Jonghee Kim, Jeffrey Willette, Sung Ju Hwang; Proceedings of the I

EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7287-7296
Dense computer vision tasks such as object detection and segmentation require effective multi-scale feature representation for detecting or classifying objects or regions with varying sizes. While Convolutional Neural Networks (CNNs) have been the dominant architectures for such tasks, recently introduced Vision Transformers (ViTs) aim to replace them as a backbone. Similar to CNNs, ViTs build a simple multi-stage structure (i.e., fine-to-coarse) for multi-scale representation with single-scale patches. In this work, with a different perspective from existing Transformers, we explore multi-scale patch embedding and multi-path structure, constructing the Multi-Path Vision Transformer (MPViT). MPViT embeds features of the same size (i.e., sequence length) with patches of different scales simultaneously by using overlapping convolutional patch embedding. Tokens of different scales are then independently fed into the Transformer encoders via multiple paths and the resulting features are aggregated, enabling both fine and coarse feature representations at the same feature level. Thanks to the diverse, multi-scale feature representations, our MPViTs scaling from tiny (5M) to base (73M) consistently achieve superior performance over state-of-the-art Vision Transformers on ImageNet classification, object detection, instance segmentation, and semantic segmentation. These extensive results demonstrate that MPViT can serve as a versatile backbone network for various vision tasks.
****
NICGSlowDown: Evaluating the Efficiency Robustness of Neural Image Caption Generation Models
Simin Chen, Zihe Song, Mirazul Haque, Cong Liu, Wei Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15365-15374
Neural image caption generation (NICG) models have received massive attention from the research community due to their excellent performance in visual understanding. Existing work focuses on improving NICG model accuracy while efficiency is less explored. However, many real-world applications require real-time feedback, which highly relies on the efficiency of NICG models. Recent research observed that the efficiency of NICG models could vary for different inputs. This observation brings in a new attack surface of NICG models, i.e., An adversary might be able to slightly change inputs to cause the NICG models to consume more computational resources. To further understand such efficiency-oriented threats, we propose a new attack approach, NICGSlowDown, to evaluate the efficiency robustness of NICG models. Our experimental results show that NICGSlowDown can generate images with human-unnoticeable perturbations that will increase the NICG model latency up to 483.86%. We hope this research could raise the community's concern about the efficiency robustness of NICG models.
****
Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation
Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11090-11100
We propose a robust and accurate method for estimating the 3D poses of two hands in close interaction from a single color image. This is a very challenging problem, as large occlusions and many confusions between the joints may happen. State-of-the-art methods solve this problem by regressing a heatmap for each joint, which requires solving two problems simultaneously: localizing the joints and recognizing them. In this work, we propose to separate these tasks by relying on a CNN to first localize joints as 2D keypoints, and on self-attention between the CNN features at these keypoints to associate them with the corresponding hand joint. The resulting architecture, which we call "Keypoint Transformer", is highly efficient as it achieves state-of-the-art performance with roughly half the number of model parameters on the InterHand2.6M dataset. We also show it can be easily extended to estimate the 3D pose of an object manipulated by one or two hands with high performance. Moreover, we created a new dataset of more than 75,000

images of two hands manipulating an object fully annotated in 3D and will make it publicly available.
****

SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing

Yichun Shi, Xiao Yang, Yangyue Wan, Xiaohui Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11254-11264

Recent studies have shown that StyleGANs provide promising prior models for downstream tasks on image synthesis and editing. However, since the latent codes of StyleGANs are designed to control global styles, it is hard to achieve a fine-grained control over synthesized images. We present SemanticStyleGAN, where a generator is trained to model local semantic parts separately and synthesizes images in a compositional way. The structure and texture of different local parts are controlled by corresponding latent codes. Experimental results demonstrate that our model provides a strong disentanglement between different spatial areas. When combined with editing methods designed for StyleGANs, it can achieve a more fine-grained control to edit synthesized or real images. The model can also be extended to other domains via transfer learning. Thus, as a generic prior model with built-in disentanglement, it could facilitate the development of GAN-based applications and enable more potential downstream tasks.
****

Accurate 3D Body Shape Regression Using Metric and Semantic Attributes

Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2718-2728

While methods that regress 3D human meshes from images have progressed rapidly, the estimated body shapes often do not capture the true human shape. This is problematic since, for many applications, accurate body shape is as important as pose. The key reason that body shape accuracy lags pose accuracy is the lack of data. While humans can label 2D joints, and these constrain 3D pose, it is not so easy to "label" 3D body shape. Since paired data with images and 3D body shape are rare, we exploit two sources of information: (1) we collect internet images of diverse "fashion" models together with a small set of anthropometric measurements; (2) we collect linguistic shape attributes for a wide range of 3D body meshes and the model images. Taken together, these datasets provide sufficient constraints to infer dense 3D shape. We exploit the anthropometric measurements and linguistic shape attributes in several novel ways to train a neural network, called SHAPY, that regresses 3D human pose and shape from an RGB image. We evaluate SHAPY on public benchmarks, but note that they either lack significant body shape variation, ground-truth shape, or clothing variation. Thus, we collect a new dataset for evaluating 3D human shape estimation, called HBW, containing photos of "Human Bodies in the Wild" for which we have ground-truth 3D body scans. On this new benchmark, SHAPY significantly outperforms state-of-the-art methods on the task of 3D body shape estimation. This is the first demonstration that 3D body shape regression from images can be trained from easy-to-obtain anthropometric measurements and linguistic shape attributes. Our model and data are available at: shapy.is.tue.mpg.de
****

VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, Vasudev Lal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21406-21415

Breakthroughs in transformer-based models have revolutionized not only the NLP field, but also vision and multimodal systems. However, although visualization and interpretability tools have become available for NLP models, internal mechanisms of vision and multimodal transformers remain largely opaque. With the success of these transformers, it is increasingly critical to understand their inner workings, as unraveling these black-boxes will lead to more capable and trustworth

y models. To contribute to this quest, we propose VL-InterpreT, which provides n ovel interactive visualizations for interpreting the attentions and hidden repre sentations in multimodal transformers. VL-InterpreT is a task agnostic and integ rated tool that (1) tracks a variety of statistics in attention heads throughout all layers for both vision and language components, (2) visualizes cross-modal and intra-modal attentions through easily readable heatmaps, and (3) plots the h idden representations of vision and language tokens as they pass through the tra nsformer layers. In this paper, we demonstrate the functionalities of VL-Interpr eT through the analysis of KD-VLP, an end-to-end pretraining vision-language mul timodal transformer-based model, in the tasks of Visual Commonsense Reasoning (V CR) and WebQA, two visual question answering benchmarks. Furthermore, we also pr esent a few interesting findings about multimodal transformer behaviors that wer e learned through our tool.

****

Label-Only Model Inversion Attacks via Boundary Repulsion

Mostafa Kahla, Si Chen, Hoang Anh Just, Ruoxi Jia; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15045-150 53

Recent studies show that the state-of-the-art deep neural networks are vulnerabl e to model inversion attacks, in which access to a model is abused to reconstruc t private training data of any given target class. Existing attacks rely on havi ng access to either the complete target model(whitebox) or the model's soft-labe ls (blackbox). However,no prior work has been done in the harder but more practi cal scenario, in which the attacker only has access to the model's predicted lab el, without a confidence measure. In this paper, we introduce an algorithm, Boun dary-Repelling Model Inversion (BREP-MI), to invert private training data using only the target model's predicted labels. The key idea of our algorithm is to ev aluate the model's predicted labels over a sphere and then estimate the directio n to reach the target class's centroid. Using the example of face recognition, w e show that the images reconstructed by BREP-MI successfully reproduce the seman tics of the private training data for various datasets and target model architec tures. We compare BREP-MI with the state-of-the-art white-box and blackbox model inversion attacks and the results show that despite assuming less knowledge abo ut the target model, BREP-MI outperforms the blackbox attack and achieves compar able results to the whitebox attack.

****

Privacy-Preserving Online AutoML for Domain-Specific Face Detection

Chenqian Yan, Yuge Zhang, Quanlu Zhang, Yaming Yang, Xinyang Jiang, Yuqing Yang, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2022, pp. 4134-4144

Despite the impressive progress of general face detection, the tuning of hyper-p arameters and architectures is still critical for the performance of a domain-sp ecific face detector. Though existing AutoML works can speedup such process, the y either require tuning from scratch for a new scenario or do not consider data privacy. To scale up, we derive a new AutoML setting from a platform perspective . In such setting, new datasets sequentially arrive at the platform, where an ar chitecture and hyper-parameter configuration is recommended to train the optimal face detector for each dataset. This, however, brings two major challenges: (1) how to predict the best configuration for any given dataset without touching th eir raw images due to the privacy concern? and (2) how to continuously improve t he AutoML algorithm from previous tasks and offer a better warm-up for future on es? We introduce "HyperFD", a new privacy-preserving online AutoML framework for face detection. At its core part, a novel meta-feature representation of a data set as well as its learning paradigm is proposed. Thanks to HyperFD, each local task (client) is able to effectively leverage the learning "experience" of previ ous tasks without uploading raw images to the platform; meanwhile, the meta-feat ure extractor is continuously learned to better trade off the bias and variance. Extensive experiments demonstrate the effectiveness and efficiency of our desig n.

****

Self-Augmented Unpaired Image Dehazing via Density and Depth Decomposition
Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2037-2046
To overcome the overfitting issue of dehazing models trained on synthetic hazy-clean image pairs, many recent methods attempted to improve models' generalization ability by training on unpaired data. Most of them simply formulate dehazing and rehazing cycles, yet ignore the physical properties of the real-world hazy environment, i.e. the haze varies with density and depth. In this paper, we propose a self-augmented image dehazing framework, termed D^4 (Dehazing via Decomposing transmission map into Density and Depth) for haze generation and removal. Instead of merely estimating transmission maps or clean content, the proposed framework focuses on exploring scattering coefficient and depth information contained in hazy and clean images. With estimated scene depth, our method is capable of re-rendering hazy images with different thicknesses which further benefits the training of the dehazing network. It is worth noting that the whole training process needs only unpaired hazy and clean images, yet succeeded in recovering the scattering coefficient, depth map and clean content from a single hazy image. Comprehensive experiments demonstrate our method outperforms state-of-the-art unpaired dehazing methods with much fewer parameters and FLOPs. Our code is available at https://github.com/YaN9-Y/D4
****
Neural 3D Video Synthesis From Multi-View Video
Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, Zhaoyang Lv; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5521-5531
We propose a novel approach for 3D video synthesis that is able to represent multi-view video recordings of a dynamic real-world scene in a compact, yet expressive representation that enables high-quality view synthesis and motion interpolation. Our approach takes the high quality and compactness of static neural radiance fields in a new direction: to a model-free, dynamic setting. At the core of our approach is a novel time-conditioned neural radiance field that represents scene dynamics using a set of compact latent codes. We are able to significantly boost the training speed and perceptual quality of the generated imagery by a novel hierarchical training scheme in combination with ray importance sampling. Our learned representation is highly compact and able to represent a 10 second 30 FPS multi-view video recording by 18 cameras with a model size of only 28MB. We demonstrate that our method can render high-fidelity wide-angle novel views at over 1K resolution, even for complex and dynamic scenes. We perform an extensive qualitative and quantitative evaluation that shows that our approach outperforms the state of the art. Project website: https://neural-3d-video.github.io/.
****
LiDAR Snowfall Simulation for Robust 3D Object Detection
Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16364-16374
3D object detection is a central task for applications such as autonomous driving, in which the system needs to localize and classify surrounding traffic agents, even in the presence of adverse weather. In this paper, we address the problem of LiDAR-based 3D object detection under snowfall. Due to the difficulty of collecting and annotating training data in this setting, we propose a physically based method to simulate the effect of snowfall on real clear-weather LiDAR point clouds. Our method samples snow particles in 2D space for each LiDAR line and uses the induced geometry to modify the measurement for each LiDAR beam accordingly. Moreover, as snowfall often causes wetness on the ground, we also simulate ground wetness on LiDAR point clouds. We use our simulation to generate partially synthetic snowy LiDAR data and leverage these data for training 3D object detection models that are robust to snowfall. We conduct an extensive evaluation using several state-of-the-art 3D object detection methods and show that our simulati

on consistently yields significant performance gains on the real snowy STF dataset compared to clear-weather baselines and competing simulation approaches, while not sacrificing performance in clear weather. Our code is available at github.com/SysCV/LiDAR_snow_sim.
****

Learning Where To Learn in Cross-View Self-Supervised Learning
Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, Toshihiko Yamasaki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14451-14460
Self-supervised learning (SSL) has made enormous progress and largely narrowed the gap with the supervised ones, where the representation learning is mainly guided by a projection into an embedding space. During the projection, current methods simply adopt uniform aggregation of pixels for embedding; however, this risks involving object-irrelevant nuisances and spatial misalignment for different augmentations. In this paper, we present a new approach, Learning Where to Learn (LEWEL), to adaptively aggregate spatial information of features, so that the projected embeddings could be exactly aligned and thus guide the feature learning better. Concretely, we reinterpret the projection head in SSL as a per-pixel projection and predict a set of spatial alignment maps from the original features by this weight-sharing projection head. A spectrum of aligned embeddings is thus obtained by aggregating the features with spatial weighting according to these alignment maps. As a result of this adaptive alignment, we observe substantial improvements on both image-level prediction and dense prediction at the same time: LEWEL improves MoCov2 by 1.6%/1.3%/0.5%/0.4% points, improves BYOL by 1.3%/1.3%/0.7%/0.6% points, on ImageNet linear/semi-supervised classification, Pascal VOC semantic segmentation, and object detection, respectively.
****

SemAffiNet: Semantic-Affine Transformation for Point Cloud Segmentation
Ziyi Wang, Yongming Rao, Xumin Yu, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11819-11829
Conventional point cloud semantic segmentation methods usually employ an encoder-decoder architecture, where mid-level features are locally aggregated to extract geometric information. However, the over-reliance on these class-agnostic local geometric representations may raise confusion between local parts from different categories that are similar in appearance or spatially adjacent. To address this issue, we argue that mid-level features can be further enhanced with semantic information, and propose semantic-affine transformation that transforms features of mid-level points belonging to different categories with class-specific affine parameters. Based on this technique, we propose SemAffiNet for point cloud semantic segmentation, which utilizes the attention mechanism in the Transformer module to implicitly and explicitly capture global structural knowledge within local parts for overall comprehension of each category. We conduct extensive experiments on the ScanNetV2 and NYUv2 datasets, and evaluate semantic-affine transformation on various 3D point cloud and 2D image segmentation baselines, where both qualitative and quantitative results demonstrate the superiority and generalization ability of our proposed approach. Code is available at https://github.com/wangzy22/SemAffiNet.
****

Sparse Object-Level Supervision for Instance Segmentation With Pixel Embeddings
Adrian Wolny, Qin Yu, Constantin Pape, Anna Kreshuk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4402-4411
Most state-of-the-art instance segmentation methods have to be trained on densely annotated images. While difficult in general, this requirement is especially daunting for biomedical images, where domain expertise is often required for annotation and no large public data collections are available for pre-training. We propose to address the dense annotation bottleneck by introducing a proposal-free segmentation approach based on non-spatial embeddings, which exploits the structure of the learned embedding space to extract individual instances in a differe

ntiable way. The segmentation loss can then be applied directly to instances and the overall pipeline can be trained in a fully- or weakly supervised manner. We consider the challenging case of positive-unlabeled supervision, where a novel self-supervised consistency loss is introduced for the unlabeled parts of the training data. We evaluate the proposed method on 2D and 3D segmentation problems in different microscopy modalities as well as on the Cityscapes and CVPPP instance segmentation benchmarks, achieving state-of-the-art results on the latter.
****

How Much More Data Do I Need? Estimating Requirements for Downstream Tasks
Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, Marc T. Law; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 275-284
Given a small training data set and a learning algorithm, how much more data is necessary to reach a target validation or test performance? This question is of critical importance in applications such as autonomous driving or medical imaging where collecting data is expensive and time-consuming. Overestimating or underestimating data requirements incurs substantial costs that could be avoided with an adequate budget. Prior work on neural scaling laws suggest that the power-law function can fit the validation performance curve and extrapolate it to larger data set sizes. We find that this does not immediately translate to the more difficult downstream task of estimating the required data set size to meet a target performance. In this work, we consider a broad class of computer vision tasks and systematically investigate a family of functions that generalize the power-law function to allow for better estimation of data requirements. Finally, we show that incorporating a tuned correction factor and collecting over multiple rounds significantly improves the performance of the data estimators. Using our guidelines, practitioners can accurately estimate data requirements of machine learning systems to gain savings in both development time and data acquisition costs.
****

Structural and Statistical Texture Knowledge Distillation for Semantic Segmentation
Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, Hongtao Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16876-16885
Existing knowledge distillation works for semantic segmentation mainly focus on transfering high-level contextual knowledge from teacher to student. However, low-level texture knowledge is also of vital importance for characterizing the local structural pattern and global statistical property, such as boundary, smoothness, regularity and color contrast, which may not be well addressed by high-level deep features. In this paper, we are intended to take full advantage of both structural and statistical texture knowledge and propose a novel Structural and Statistical Texture Knowledge Distillation (SSTKD) framework for Semantic Segmentation. Specifically, for structural texture knowledge, we introduce a Contourlet Decomposition Module (CDM) that decomposes low-level features with iterative laplacian pyramid and directional filter bank to mine the structural texture knowledge. For statistical knowledge, we propose a Denoised Texture Intensity Equalization Module (DTIEM) to adaptively extract and enhance statistical texture knowledge through heuristics iterative quantization and denoised operation. Finally, each knowledge learning is supervised by an individual loss function, forcing the student network to mimic the teacher better from a broader perspective. Experiments show that the proposed method achieves state-of-the-art performance on Cityscapes, Pascal VOC 2012 and ADE20K datasets.
****

Shapley-NAS: Discovering Operation Contribution for Neural Architecture Search
Han Xiao, Ziwei Wang, Zheng Zhu, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11892-11901
In this paper, we propose a Shapley value based method to evaluate operation contribution (Shapley-NAS) for neural architecture search. Differentiable architecture search (DARTS) acquires the optimal architectures by optimizing the architec

ture parameters with gradient descent, which significantly reduces the search cost. However, the magnitude of architecture parameters updated by gradient descent fails to reveal the actual operation importance to the task performance and therefore harms the effectiveness of obtained architectures. By contrast, we propose to evaluate the direct influence of operations on validation accuracy. To deal with the complex relationships between supernet components, we leverage Shapley value to quantify their marginal contributions by considering all possible combinations. Specifically, we iteratively optimize the supernet weights and update the architecture parameters by evaluating operation contributions via Shapley value, so that the optimal architectures are derived by selecting the operations that contribute significantly to the tasks. Since the exact computation of Shapley value is NP-hard, the Monte-Carlo sampling based algorithm with early truncation is employed for efficient approximation, and the momentum update mechanism is adopted to alleviate fluctuation of the sampling process. Extensive experiments on various datasets and various search spaces show that our Shapley-NAS outperforms the state-of-the-art methods by a considerable margin with light search cost. The code is available at https://github.com/Euphoria16/Shapley-NAS.git.

****

The Implicit Values of a Good Hand Shake: Handheld Multi-Frame Neural Depth Refinement

Ilya Chugunov, Yuxuan Zhang, Zhihao Xia, Xuaner Zhang, Jiawen Chen, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2852-2862

Modern smartphones can continuously stream multi-megapixel RGB images at 60Hz, synchronized with high-quality 3D pose information and low-resolution LiDAR-driven depth estimates. During a snapshot photograph, the natural unsteadiness of the photographer's hands offers millimeter-scale variation in camera pose, which we can capture along with RGB and depth in a circular buffer. In this work we explore how, from a bundle of these measurements acquired during viewfinding, we can combine dense micro-baseline parallax cues with kilopixel LiDAR depth to distill a high-fidelity depth map. We take a test-time optimization approach and train a coordinate MLP to output photometrically and geometrically consistent depth estimates at the continuous coordinates along the path traced by the photographer's natural hand shake. With no additional hardware, artificial hand motion, or user interaction beyond the press of a button, our proposed method brings high-resolution depth estimates to point-and-shoot "tabletop" photography -- textured objects at close range.

****

Learning What Not To Segment: A New Perspective on Few-Shot Segmentation

Chunbo Lang, Gong Cheng, Binfei Tu, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8057-8067

Recently few-shot segmentation (FSS) has been extensively developed. Most previous works strive to achieve generalization through the meta-learning framework derived from classification tasks; however, the trained models are biased towards the seen classes instead of being ideally class-agnostic, thus hindering the recognition of new concepts. This paper proposes a fresh and straightforward insight to alleviate the problem. Specifically, we apply an additional branch (base learner) to the conventional FSS model (meta learner) to explicitly identify the targets of base classes, i.e., the regions that do not need to be segmented. Then, the coarse results output by these two learners in parallel are adaptively integrated to yield precise segmentation prediction. Considering the sensitivity of meta learner, we further introduce an adjustment factor to estimate the scene differences between the input image pairs for facilitating the model ensemble forecasting. The substantial performance gains on PASCAL-5i and COCO-20i verify the effectiveness, and surprisingly, our versatile scheme sets a new state-of-the-art even with two plain learners. Moreover, in light of the unique nature of the proposed approach, we also extend it to a more realistic but challenging setting, i.e., generalized FSS, where the pixels of both base and novel classes are required to be determined. The source code is available at github.com/chunbolang/BAM.

****
Blended Diffusion for Text-Driven Editing of Natural Images
Omri Avrahami, Dani Lischinski, Ohad Fried; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18208-18218
Natural language offers a highly intuitive interface for image editing. In this paper, we introduce the first solution for performing local (region-based) edits in generic natural images, based on a natural language description along with an ROI mask. We achieve our goal by leveraging and combining a pretrained language-image model (CLIP), to steer the edit towards a user-provided text prompt, with a denoising diffusion probabilistic model (DDPM) to generate natural-looking results. To seamlessly fuse the edited region with the unchanged parts of the image, we spatially blend noised versions of the input image with the local text-guided diffusion latent at a progression of noise levels. In addition, we show that adding augmentations to the diffusion process mitigates adversarial results. We compare against several baselines and related methods, both qualitatively and quantitatively, and show that our method outperforms these solutions in terms of overall realism, ability to preserve the background and matching the text. Finally, we show several text-driven editing applications, including adding a new object to an image, removing/replacing/altering existing objects, background replacement, and image extrapolation.
****
Towards Unsupervised Domain Generalization
Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, Haoxin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4910-4920
Domain generalization (DG) aims to help models trained on a set of source domains generalize better on unseen target domains. The performances of current DG methods largely rely on sufficient labeled data, which are usually costly or unavailable, however. Since unlabeled data are far more accessible, we seek to explore how unsupervised learning can help deep models generalize across domains. Specifically, we study a novel generalization problem called unsupervised domain generalization (UDG), which aims to learn generalizable models with unlabeled data and analyze the effects of pre-training on DG. In UDG, models are pretrained with unlabeled data from various source domains before being trained on labeled source data and eventually tested on unseen target domains. Then we propose a method named Domain-Aware Representation LearnING (DARLING) to cope with the significant and misleading heterogeneity within unlabeled pretraining data and severe distribution shifts between source and target data. Surprisingly we observe that DARLING can not only counterbalance the scarcity of labeled data but also further strengthen the generalization ability of models when the labeled data are insufficient. As a pretraining approach, DARLING shows superior or comparable performance compared with ImageNet pretraining protocol even when the available data are unlabeled and of a vastly smaller amount compared to ImageNet, which may shed light on improving generalization with large-scale unlabeled data.
****
HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening
Wele Gedara Chaminda Bandara, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1767-1777
Pansharpening aims to fuse a registered high-resolution panchromatic image (PAN) with a low-resolution hyperspectral image (LR-HSI) to generate an enhanced HSI with high spectral and spatial resolution. Existing pansharpening approaches neglect using an attention mechanism to transfer HR texture features from PAN to LR-HSI features, resulting in spatial and spectral distortions. In this paper, we present a novel attention mechanism for pansharpening called HyperTransformer, in which features of LR-HSI and PAN are formulated as queries and keys in a transformer, respectively. HyperTransformer consists of three main modules, namely two separate feature extractors for PAN and HSI, a multi-head feature soft attention module, and a spatial-spectral feature fusion module. Such a network improves both spatial and spectral quality measures of the pansharpened HSI by learning

cross-feature space dependencies and long-range details of PAN and LR-HSI. Furth
ermore, HyperTransformer can be utilized across multiple spatial scales at the b
ackbone for obtaining improved performance. Extensive experiments conducted on t
hree widely used datasets demonstrate that HyperTransformer achieves significant
 improvement over the state-of-the-art methods on both spatial and spectral qual
ity measures. Implementation code and pre-trained weights can be accessed at htt
ps://github.com/wgcban/HyperTransformer.
****

Segment-Fusion: Hierarchical Context Fusion for Robust 3D Semantic Segmentation
Anirud Thyagharajan, Benjamin Ummenhofer, Prashant Laddha, Om Ji Omer, Sreenivas
 Subramoney; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2022, pp. 1236-1245
3D semantic segmentation is a fundamental building block for several scene under
standing applications such as autonomous driving, robotics and AR/VR. Several st
ate-of-the-art semantic segmentation models suffer from the part-misclassificati
on problem, wherein parts of the same object are labelled incorrectly. Previous
methods have utilized hierarchical, iterative methods to fuse semantic and insta
nce information, but they lack learnability in context fusion, and are computati
onally complex and heuristic driven. This paper presents Segment-Fusion, a novel
 attention-based method for hierarchical fusion of semantic and instance informa
tion to address the part misclassifications. The presented method includes a gra
ph segmentation algorithm for grouping points into segments that pools point-wis
e features into segment-wise features, a learnable attention-based network to fu
se these segments based on their semantic and instance features, and followed by
 a simple yet effective connected component labelling algorithm to convert segme
nt features to instance labels. Segment-Fusion can be flexibly employed with any
 network architecture for semantic/instance segmentation. It improves the qualit
ative and quantitative performance of several semantic segmentation backbones by
 upto 5% on the ScanNet and S3DIS datasets.
****

Robust Invertible Image Steganography
Youmin Xu, Chong Mou, Yujie Hu, Jingfen Xie, Jian Zhang; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 787
5-7884
Image steganography aims to hide secret images into a container image, where the
 secret is hidden from human vision and can be restored when necessary. Previous
 image steganography methods are limited in hiding capacity and robustness, comm
only vulnerable to distortion on container images such as Gaussian noise, Poisso
n noise, and lossy compression. This paper presents a novel flow-based framework
 for robust invertible image steganography, dubbed as RIIS. We introduce the con
ditional normalizing flow to model the distribution of the redundant high-freque
ncy component with the condition of the container image. Moreover, a well-design
ed container enhancement module (CEM) also contributes to the robust reconstruct
ion. To regulate the network parameters for different distortion levels, we prop
ose a distortion-guided modulation (DGM) over flow-based blocks to make it a one
-size-fits-all model. In terms of both clean and distorted image steganography,
extensive experiments reveal that the proposed RIIS efficiently improves the rob
ustness while maintaining imperceptibility and capacity. As far as we know, we a
re the first learning-based scheme to enhance the robustness of image steganogra
phy in the literature. The guarantee of steganography robustness significantly b
roadens the application of steganography in real-world applications.
****

Entropy-Based Active Learning for Object Detection With Progressive Diversity Co
nstraint
Jiaxi Wu, Jiaxin Chen, Di Huang; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2022, pp. 9397-9406
Active learning is a promising alternative to alleviate the issue of high annota
tion cost in the computer vision tasks by consciously selecting more informative
 samples to label. Active learning for object detection is more challenging and
existing efforts on it are relatively rare. In this paper, we propose a novel hy

brid approach to address this problem, where the instance-level uncertainty and diversity are jointly considered in a bottom-up manner. To balance the computational complexity, the proposed approach is designed as a two-stage procedure. At the first stage, an Entropy-based Non-Maximum Suppression (ENMS) is presented to estimate the uncertainty of every image, which performs NMS according to the entropy in the feature space to remove predictions with redundant information gains. At the second stage, a diverse prototype (DivProto) strategy is explored to ensure the diversity across images by progressively converting it into the intra-class and inter-class diversities of the entropy-based class-specific prototypes. Extensive experiments are conducted on MS COCO and Pascal VOC, and the proposed approach achieves state of the art results and significantly outperforms the other counterparts, highlighting its superiority.

****

BE-STI: Spatial-Temporal Integrated Network for Class-Agnostic Motion Prediction With Bidirectional Enhancement

Yunlong Wang, Hongyu Pan, Jun Zhu, Yu-Huan Wu, Xin Zhan, Kun Jiang, Diange Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17093-17102

Determining the motion behavior of inexhaustible categories of traffic participants is critical for autonomous driving. In recent years, there has been a rising concern in performing class-agnostic motion prediction directly from the captured sensor data, like LiDAR point clouds or the combination of point clouds and images. Current motion prediction frameworks tend to perform joint semantic segmentation and motion prediction and face the trade-off between the performance of these two tasks. In this paper, we propose a novel Spatial-Temporal Integrated network with Bidirectional Enhancement, BE-STI, to improve the temporal motion prediction performance by spatial semantic features, which points out an efficient way to combine semantic segmentation and motion prediction. Specifically, we propose to enhance the spatial features of each individual point cloud with the similarity among temporal neighboring frames and enhance the global temporal features with the spatial difference among non-adjacent frames in a coarse-to-fine fashion. Extensive experiments on nuScenes and Waymo Open Dataset show that our proposed framework outperforms all state-of-the-art LiDAR-based and RGB+LiDAR-based methods with remarkable margins by using only point clouds as input.

****

A Structured Dictionary Perspective on Implicit Neural Representations

Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, Pascal Frossard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19228-19238

Implicit neural representations (INRs) have recently emerged as a promising alternative to classical discretized representations of signals. Nevertheless, despite their practical success, we still do not understand how INRs represent signals. We propose a novel unified perspective to theoretically analyse INRs. Leveraging results from harmonic analysis and deep learning theory, we show that most INR families are analogous to structured signal dictionaries whose atoms are integer harmonics of the set of initial mapping frequencies. This structure allows INRs to express signals with an exponentially increasing frequency support using a number of parameters that only grows linearly with depth. We also explore the inductive bias of INRs exploiting recent results about the empirical neural tangent kernel (NTK). Specifically, we show that the eigenfunctions of the NTK can be seen as dictionary atoms whose inner product with the target signal determines the final performance of their reconstruction. In this regard, we reveal that meta-learning has a reshaping effect on the NTK analogous to dictionary learning, building dictionary atoms as a combination of the examples seen during meta-training. Our results permit to design and tune novel INR architectures, but can also be of interest for the wider deep learning theory community.

****

Egocentric Deep Multi-Channel Audio-Visual Active Speaker Localization

Hao Jiang, Calvin Murdock, Vamsi Krishna Ithapu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10544-10552

Augmented reality devices have the potential to enhance human perception and ena
ble other assistive functionalities in complex conversational environments. Effe
ctively capturing the audio-visual context necessary for understanding these soc
ial interactions first requires detecting and localizing the voice activities of
 the device wearer and the surrounding people. These tasks are challenging due t
o their egocentric nature: the wearer's head motion may cause motion blur, surro
unding people may appear in difficult viewing angles, and there may be occlusion
s, visual clutter, audio noise, and bad lighting. Under these conditions, previo
us state-of-the-art active speaker detection methods do not give satisfactory re
sults. Instead, we tackle the problem from a new setting using both video and mu
lti-channel microphone array audio. We propose a novel end-to-end deep learning
 approach that is able to give robust voice activity detection and localization r
esults. In contrast to previous methods, our method localizes active speakers fr
om all possible directions on the sphere, even outside the camera's field of vie
w, while simultaneously detecting the device wearer's own voice activity. Our ex
periments show that the proposed method gives superior results, can run in real
time, and is robust against noise and clutter.
****

Vision-Language Pre-Training With Triple Contrastive Learning
Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zen
g, Trishul Chilimbi, Junzhou Huang; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 15671-15680
Vision-language representation learning largely benefits from image-text alignme
nt through contrastive losses (e.g., InfoNCE loss). The success of this alignmen
t strategy is attributed to its capability in maximizing the mutual information
(MI) between an image and its matched text. However, simply performing cross-mod
al alignment (CMA) ignores data potential within each modality, which may result
 in degraded representations. For instance, although CMA-based models are able t
o map image-text pairs close together in the embedding space, they fail to ensur
e that similar inputs from the same modality stay close by. This problem can get
 even worse when the pre-training data is noisy. In this paper, we propose tripl
e contrastive learning (TCL) for vision-language pre-training by leveraging both
 cross-modal and intra-modal self-supervision. Besides CMA, TCL introduces an in
tra-modal contrastive objective to provide complementary benefits in representat
ion learning. To take advantage of localized and structural information from ima
ge and text input, TCL further maximizes the average MI between local regions of
 image/text and their global summary. To the best of our knowledge, ours is the
first work that takes into account local structure information for multi-modalit
y representation learning. Experimental evaluations show that our approach is co
mpetitive and achieves the new state of the art on various common down-stream vi
sion-language tasks such as image-text retrieval and visual question answering.
****

Structure-Aware Flow Generation for Human Body Reshaping
Jianqiang Ren, Yuan Yao, Biwen Lei, Miaomiao Cui, Xuansong Xie; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,
pp. 7754-7763
Body reshaping is an important procedure in portrait photo retouching. Due to th
e complicated structure and multifarious appearance of human bodies, existing me
thods either fall back on the 3D domain via body morphable model or resort to ke
ypoint-based image deformation, leading to inefficiency and unsatisfied visual q
uality. In this paper, we address these limitations by formulating an end-to-end
 flow generation architecture under the guidance of body structural priors, incl
uding skeletons and Part Affinity Fields, and achieve unprecedentedly controllab
le performance under arbitrary poses and garments. A compositional attention mec
hanism is introduced for capturing both visual perceptual correlations and struc
tural associations of the human body to reinforce the manipulation consistency a
mong related parts. For a comprehensive evaluation, we construct the first large
-scale body reshaping dataset, namely BR-5K, which contains 5,000 portrait photo
s as well as professionally retouched targets. Extensive experiments demonstrate
 that our approach significantly outperforms existing state-of-the-art methods i

n terms of visual performance, controllability, and efficiency. The dataset is a vailable at our website: https://github.com/JianqiangRen/FlowBasedBodyReshaping.
****

Practical Learned Lossless JPEG Recompression With Multi-Level Cross-Channel Ent ropy Model in the DCT Domain
Lina Guo, Xinjie Shi, Dailan He, Yuanyuan Wang, Rui Ma, Hongwei Qin, Yan Wang; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5862-5871
JPEG is a popular image compression method widely used by individuals, data cent er, cloud storage and network filesystems. However, most recent progress on imag e compression mainly focuses on uncompressed images while ignoring trillions of already-existing JPEG images. To compress these JPEG images adequately and resto re them back to JPEG format losslessly when needed, we propose a deep learning b ased JPEG recompression method that operates on DCT domain and propose a Multi-L evel Cross-Channel Entropy Model to compress the most informative Y component. E xperiments show that our method achieves state-of-the-art performance compared w ith traditional JPEG recompression methods including Lepton, JPEG XL and CMIX. T o the best of our knowledge, this is the first learned compression method that l osslessly transcodes JPEG images to more storage-saving bitstreams.
****

Fourier PlenOctrees for Dynamic Radiance Field Rendering in Real-Time
Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zha ng, Minye Wu, Jingyi Yu, Lan Xu; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2022, pp. 13524-13534
Implicit neural representations such as Neural Radiance Field (NeRF) have focuse d mainly on modeling static objects captured under multi-view settings where rea l-time rendering can be achieved with smart data structures, e.g., PlenOctree. I n this paper, we present a novel Fourier PlenOctree (FPO) technique to tackle ef ficient neural modeling and real-time rendering of dynamic scenes captured under the free-view video (FVV) setting. The key idea in our FPO is a novel combinati on of generalized NeRF, PlenOctree representation, volumetric fusion and Fourier transform. To accelerate FPO construction, we present a novel coarse-to-fine fu sion scheme that leverages the generalizable NeRF technique to generate the tree via spatial blending. To tackle dynamic scenes, we tailor the implicit network to model the Fourier coefficients of time-varying density and color attributes. Finally, we construct the FPO and train the Fourier coefficients directly on the leaves of a union PlenOctree structure of the dynamic sequence. We show that th e resulting FPO enables compact memory overload to handle dynamic objects and su pports efficient fine-tuning. Extensive experiments show that the proposed metho d is 3000 times faster than the original NeRF and achieves over an order of magn itude acceleration over SOTA while preserving high visual quality for the free-v iewpoint rendering of unseen dynamic scenes.
****

Learning To Answer Questions in Dynamic Audio-Visual Scenarios
Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, Di Hu; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19108-19118
In this paper, we focus on the Audio-Visual Question Answering (AVQA) task, whic h aims to answer questions regarding different visual objects, sounds, and their associations in videos. The problem requires comprehensive multimodal understan ding and spatio-temporal reasoning over audio-visual scenes. To benchmark this t ask and facilitate our study, we introduce a large-scale MUSIC-AVQA dataset, whi ch contains more than 45K question-answer pairs covering 33 different question t emplates spanning over different modalities and question types. We develop sever al baselines and introduce a spatio-temporal grounded audio-visual network for t he AVQA problem. Our results demonstrate that AVQA benefits from multisensory pe rception and our model outperforms recent A-, V-, and AVQA approaches. We believ e that our built dataset has the potential to serve as testbed for evaluating an d promoting progress in audio-visual scene understanding and spatio-temporal rea soning. Code and dataset: http://gewu-lab.github.io/MUSIC-AVQA/

****
Leveraging Equivariant Features for Absolute Pose Regression
Mohamed Adel Musallam, Vincent Gaudillière, Miguel Ortiz del Castillo, Kassem Al Ismaeil, Djamila Aouada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6876-6886
While end-to-end approaches have achieved state-of-the-art performance in many perception tasks, they are not yet able to compete with 3D geometry-based methods in pose estimation. Moreover, absolute pose regression has been shown to be more related to image retrieval. As a result, we hypothesize that the statistical features learned by classical Convolutional Neural Networks do not carry enough geometric information to reliably solve this inherently geometric task. In this paper, we demonstrate how a translation and rotation equivariant Convolutional Neural Network directly induces representations of camera motions into the feature space. We then show that this geometric property allows for implicitly augmenting the training data under a whole group of image plane-preserving transformations. Therefore, we argue that directly learning equivariant features is preferable than learning data-intensive intermediate representations. Comprehensive experimental validation demonstrates that our lightweight model outperforms existing ones on standard datasets.
****
Synthetic Aperture Imaging With Events and Frames
Wei Liao, Xiang Zhang, Lei Yu, Shijie Lin, Wen Yang, Ning Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17735-17744
The Event-based Synthetic Aperture Imaging (E-SAI) has recently been proposed to see through extremely dense occlusions. However, the performance of E-SAI is not consistent under sparse occlusions due to the dramatic decrease of signal events. This paper addresses this problem by leveraging the merits of both events and frames, leading to a fusion-based SAI (EF-SAI) that performs consistently under the different densities of occlusions. In particular, we first extract the feature from events and frames via multi-modal feature encoders and then apply a multi-stage fusion network for cross-modal enhancement and density-aware feature selection. Finally, a CNN decoder is employed to generate occlusion-free visual images from selected features. Extensive experiments show that our method effectively tackles varying densities of occlusions and achieves superior performance to the state-of-the-art SAI methods. Codes and datasets are available at https://github.com/smjsc/EF-SAI
****
CLIP-Event: Connecting Text and Images With Event Structures
Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16420-16429
Vision-language (V+L) pretraining models have achieved great success in supporting multimedia applications by understanding the alignments between images and text. While existing vision-language pretraining models primarily focus on understanding objects in images or entities in text, they often ignore the alignment at the level of events and their argument structures. In this work, we propose a contrastive learning framework to enforce vision-language pretraining models to comprehend events and associated argument (participant) roles. To achieve this, we take advantage of text information extraction technologies to obtain event structural knowledge, and utilize multiple prompt functions to contrast difficult negative descriptions by manipulating event structures. We also design an event graph alignment loss based on optimal transport to capture event argument structures. In addition, we collect a large event-rich dataset (106,875 images) for pretraining, which provides a more challenging image retrieval benchmark to assess the understanding of complicated lengthy sentences. Experiments show that our zero-shot CLIP-Event outperforms the state-of-the-art supervised model in argument extraction on Multimedia Event Extraction, achieving more than 5% absolute F-score gain in event extraction, as well as significant improvements on a variety of downstream tasks under zero-shot settings.

****
MonoGround: Detecting Monocular 3D Objects From the Ground
Zequn Qin, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3793-3802
Monocular 3D object detection has attracted great attention for its advantages in simplicity and cost. Due to the ill-posed 2D to 3D mapping essence from the monocular imaging process, monocular 3D object detection suffers from inaccurate depth estimation and thus has poor 3D detection results. To alleviate this problem, we propose to introduce the ground plane as a prior in the monocular 3d object detection. The ground plane prior serves as an additional geometric condition to the ill-posed mapping and an extra source in depth estimation. In this way, we can get a more accurate depth estimation from the ground. Meanwhile, to take full advantage of the ground plane prior, we propose a depth-align training strategy and a precise two-stage depth inference method tailored for the ground plane prior. It is worth noting that the introduced ground plane prior requires no extra data sources like LiDAR, stereo images, and depth information. Extensive experiments on the KITTI benchmark show that our method could achieve state-of-the-art results compared with other methods while maintaining a very fast speed. Our code, models, and training logs are available at https://github.com/cfzd/MonoGround.
****
Deep Visual Geo-Localization Benchmark
Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, Barbara Caputo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5396-5407
In this paper, we propose a new open-source benchmarking framework for Visual Geo-localization (VG) that allows to build, train, and test a wide range of commonly used architectures, with the flexibility to change individual components of a geo-localization pipeline. The purpose of this framework is twofold: i) gaining insights into how different components and design choices in a VG pipeline impact the final results, both in terms of performance (recall@N metric) and system requirements (such as execution time and memory consumption); ii) establish a systematic evaluation protocol for comparing different methods. Using the proposed framework, we perform a large suite of experiments which provide criteria for choosing backbone, aggregation and negative mining depending on the use-case and requirements. We also assess the impact of engineering techniques like pre/post-processing, data augmentation and image resizing, showing that better performance can be obtained through somewhat simple procedures: for example, downscaling the images' resolution to 80% can lead to similar results with a 36% savings in extraction time and dataset storage requirement. Code and trained models are available at https://deep-vg-bench.herokuapp.com/.
****
Scaling Up Vision-Language Pre-Training for Image Captioning
Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, Lijuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17980-17989
In recent years, we have witnessed significant performance boost in the image captioning task based on vision-language pre-training (VLP). Scale is believed to be an important factor for this advance. However, most existing work only focuses on pre-training transformers with moderate sizes (e.g., 12 or 24 layers) on roughly 4 million images. In this paper, we present LEMON, a LargE-scale iMage captiONer, and provide the first empirical study on the scaling behavior of VLP for image captioning. We use the state-of-the-art VinVL model as our reference model, which consists of an image feature extractor and a transformer model, and scale the transformer both up and down, with model sizes ranging from 13 to 675 million parameters. In terms of data, we conduct experiments with up to 200 million image-text pairs which are automatically collected from web based on the alt attribute of the image (dubbed as ALT200M). Extensive analysis helps to characterize the performance trend as the model size and the pre-training data size increase. We also compare different training recipes, especially for training on large

-scale noisy data. As a result, LEMON achieves new state of the arts on several major image captioning benchmarks, including COCO Caption, nocaps, and Conceptual Captions. We also show LEMON can generate captions with long-tail visual concepts when used in a zero-shot manner.

****

Semiconductor Defect Detection by Hybrid Classical-Quantum Deep Learning

Yuan-Fu Yang, Min Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2323-2332

With the rapid development of artificial intelligence and autonomous driving technology, the demand for semiconductors is projected to rise substantially. However, the massive expansion of semiconductor manufacturing and the development of new technology will bring many defect wafers. If these defect wafers have not been correctly inspected, the ineffective semiconductor processing on these defect wafers will cause additional impact to our environment, such as excessive carbon dioxide emission and energy consumption. In this paper, we utilize the information processing advantages of quantum computing to promote the defect learning defect review (DLDR). We propose a classical-quantum hybrid algorithm for deep learning on near-term quantum processors. By tuning parameters implemented on it, quantum circuit driven by our framework learns a given DLDR task, include of wafer defect map classification, defect pattern classification, and hotspot detection. In addition, we explore parametrized quantum circuits with different expressibility and entangling capacities. These results can be used to build a future roadmap to develop circuit-based quantum deep learning for semiconductor defect detection.

****

StyleGAN-V: A Continuous Video Generator With the Price, Image Quality and Perks of StyleGAN2

Ivan Skorokhodov, Sergey Tulyakov, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3626-3636

Videos show continuous events, yet most -- if not all -- video synthesis frameworks treat them discretely in time. In this work, we think of videos of what they should be -- time-continuous signals, and extend the paradigm of neural representations to build a continuous-time video generator. For this, we first design continuous motion representations through the lens of positional embeddings. Then, we explore the question of training on very sparse videos and demonstrate that a good generator can be learned by using as few as 2 frames per clip. After that, we rethink the traditional image + video discriminators pair and design a holistic discriminator that aggregates temporal information by simply concatenating frames' features. This decreases the training cost and provides richer learning signal to the generator, making it possible to train directly on 1024x1024 videos for the first time. We build our model on top of StyleGAN2 and it is just 5% more expensive to train at the same resolution while achieving almost the same image quality. Moreover, our latent space features similar properties, enabling spatial manipulations that our method can propagate in time. We can generate arbitrarily long videos at arbitrary high frame rate, while prior work struggles to generate even 64 frames at a fixed rate. Our model is tested on four modern 256x256 and one 1024x1024-resolution video synthesis benchmarks. In terms of sheer metrics, it performs on average 30% better than the closest runner-up. Project website: https://universome.github.io/stylegan-v.

****

Towards Practical Deployment-Stage Backdoor Attack on Deep Neural Networks

Xiangyu Qi, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, Kai Bu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13347-13357

One major goal of the AI security community is to securely and reliably produce and deploy deep learning models for real-world applications. To this end, data poisoning based backdoor attacks on deep neural networks (DNNs) in the production stage (or training stage) and corresponding defenses are extensively explored in recent years. Ironically, backdoor attacks in the deployment stage, which can

often happen in unprofessional users' devices and are thus arguably far more thr
eatening in real-world scenarios, draw much less attention of the community. We
attribute this imbalance of vigilance to the weak practicality of existing deplo
yment-stage backdoor attack algorithms and the insufficiency of real-world attac
k demonstrations. To fill the blank, in this work, we study the realistic threat
 of deployment-stage backdoor attacks on DNNs. We base our study on a commonly u
sed deployment-stage attack paradigm --- adversarial weight attack, where advers
aries selectively modify model weights to embed backdoor into deployed DNNs. To
approach realistic practicality, we propose the first gray-box and physically re
alizable weights attack algorithm for backdoor injection, namely subnet replacem
ent attack (SRA), which only requires architecture information of the victim mod
el and can support physical triggers in the real world. Extensive experimental s
imulations and system-level real-world attack demonstrations are conducted. Our
results not only suggest the effectiveness and practicality of the proposed atta
ck algorithm, but also reveal the practical risk of a novel type of computer vir
us that may widely spread and stealthily inject backdoor into DNN models in user
 devices. By our study, we call for more attention to the vulnerability of DNNs
in the deployment stage.
****

Scaling Vision Transformers

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, Lucas Beyer; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p
p. 12104-12113

Attention-based neural networks such as the Vision Transformer (ViT) have recent
ly attained state-of-the-art results on many computer vision benchmarks. Scale i
s a primary ingredient in attaining excellent results, therefore, understanding
a model's scaling properties is a key to designing future generations effectivel
y. While the laws for scaling Transformer language models have been studied, it
is unknown how Vision Transformers scale. To address this, we scale ViT models a
nd data, both up and down, and characterize the relationships between error rate
, data, and compute. Along the way, we refine the architecture and training of V
iT, reducing memory consumption and increasing accuracy of the resulting models.
 As a result, we successfully train a ViT model with two billion parameters, whi
ch attains a new state-of-the-art on ImageNet of 90.45% top-1 accuracy. The mode
l also performs well for few-shot transfer, for example, reaching 84.86% top-1 a
ccuracy on ImageNet with only 10 examples per class.
****

Unsupervised Action Segmentation by Joint Representation Learning and Online Clu
stering

Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M. Zeeshan Zia, Quoc-Hu
y Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2022, pp. 20174-20185

We present a novel approach for unsupervised activity segmentation which uses vi
deo frame clustering as a pretext task and simultaneously performs representatio
n learning and online clustering. This is in contrast with prior works where rep
resentation learning and clustering are often performed sequentially. We leverag
e temporal information in videos by employing temporal optimal transport. In par
ticular, we incorporate a temporal regularization term which preserves the tempo
ral order of the activity into the standard optimal transport module for computi
ng pseudo-label cluster assignments. The temporal optimal transport module enabl
es our approach to learn effective representations for unsupervised activity seg
mentation. Furthermore, previous methods require storing learned features for th
e entire dataset before clustering them in an offline manner, whereas our approa
ch processes one mini-batch at a time in an online manner. Extensive evaluations
 on three public datasets, i.e. 50-Salads, YouTube Instructions, and Breakfast,
and our dataset, i.e., Desktop Assembly, show that our approach performs on par
with or better than previous methods, despite having significantly less memory c
onstraints.
****

Pin the Memory: Learning To Generalize Semantic Segmentation

Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, Kwanghoon Sohn; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, p p. 4350-4360

The rise of deep neural networks has led to several breakthroughs for semantic s egmentation. In spite of this, a model trained on source domain often fails to w ork properly in new challenging domains, that is directly concerned with the gen eralization capability of the model. In this paper, we present a novel memory-gu ided domain generalization method for semantic segmentation based on meta-learni ng framework. Especially, our method abstracts the conceptual knowledge of seman tic classes into categorical memory which is constant beyond the domains. Upon t he meta-learning concept, we repeatedly train memory-guided networks and simulat e virtual test to 1) learn how to memorize a domain-agnostic and distinct inform ation of classes and 2) offer an externally settled memory as a class-guidance t o reduce the ambiguity of representation in the test data of arbitrary unseen do main. To this end, we also propose memory divergence and feature cohesion losses , which encourage to learn memory reading and update processes for category-awar e domain generalization. Extensive experiments for semantic segmentation demonst rate the superior generalization capability of our method over state-of-the-art works on various benchmarks.

****

LISA: Learning Implicit Shape and Appearance of Hands

Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richa rd Newcombe, Lingni Ma; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2022, pp. 20533-20543

This paper proposes a do-it-all neural model of human hands, named LISA. The mod el can capture accurate hand shape and appearance, generalize to arbitrary hand subjects, provide dense surface correspondences, be reconstructed from images in the wild and easily animated. We train LISA by minimizing the shape and appeara nce losses on a large set of multi-view RGB image sequences annotated with coars e 3D poses of the hand skeleton. For a 3D point in the hand local coordinate, ou r model predicts the color and the signed distance with respect to each hand bon e independently, and then combines the per-bone predictions using predicted skin ning weights. The shape, color and pose representations are disentangled by desi gn, allowing to estimate or animate only selected parameters. We experimentally demonstrate that LISA can accurately reconstruct a dynamic hand from monocular o r multi-view sequences, achieving a noticeably higher quality of reconstructed h and shapes compared to baseline approaches. Project page: https://www.iri.upc.ed u/people/ecorona/lisa/.

****

DiGS: Divergence Guided Shape Implicit Neural Representation for Unoriented Poin t Clouds

Yizhak Ben-Shabat, Chamin Hewa Koneputugodage, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp . 19323-19332

Shape implicit neural representations (INR) have recently shown to be effective in shape analysis and reconstruction tasks. Existing INRs require point coordina tes to learn the implicit level sets of the shape. When a normal vector is avail able for each point, a higher fidelity representation can be learned, however no rmal vectors are often not provided as raw data. Furthermore, the method's initi alization has been shown to play a crucial role for surface reconstruction. In t his paper, we propose a divergence guided shape representation learning approach that does not require normal vectors as input. We show that incorporating a sof t constraint on the divergence of the distance function favours smooth solutions that reliably orients gradients to match the unknown normal at each point, in s ome cases even better than approaches that use ground truth normal vectors direc tly. Additionally, we introduce a novel geometric initialization method for sinu soidal INRs that further improves convergence to the desired solution. We evalua te the effectiveness of our approach on the task of surface reconstruction and s hape space learning and show SOTA performance compared to other unoriented metho ds.

****
Iterative Deep Homography Estimation
Si-Yuan Cao, Jianxin Hu, Zehua Sheng, Hui-Liang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1879-1888
We propose Iterative Homography Network, namely IHN, a new deep homography estimation architecture. Different from previous works that achieve iterative refinement by network cascading or untrainable IC-LK iterator, the iterator of IHN has tied weights and is completely trainable. IHN achieves state-of-the-art accuracy on several datasets including challenging scenes. We propose 2 versions of IHN: (1) IHN for static scenes, (2) IHN-mov for dynamic scenes with moving objects. Both versions can be arranged in 1-scale for efficiency or 2-scale for accuracy. We show that the basic 1-scale IHN already outperforms most of the existing methods. On a variety of datasets, the 2-scale IHN outperforms all competitors by a large gap. We introduce IHN-mov by producing an inlier mask to further improve the estimation accuracy of moving-objects scenes. We experimentally show that the iterative framework of IHN can achieve 95% error reduction while considerably saving network parameters. When processing sequential image pairs, IHN can achieve 32.7 fps, which is about 8x the speed of IC-LK iterator. Source code is available at https://github.com/imdumpl78/IHN.
****
Semi-Supervised Learning of Semantic Correspondence With Pseudo-Labels
Jiwon Kim, Kwangrok Ryoo, Junyoung Seo, Gyuseong Lee, Daehwan Kim, Hansang Cho, Seungryong Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19699-19709
Establishing dense correspondences across semantically similar images remains a challenging task due to the significant intra-class variations and background clutters. Traditionally, a supervised loss was used for training the matching networks, which requires tremendous manually-labeled data, while some methods suggested a self-supervised or weakly-supervised loss to mitigate the reliance on the labeled data, but with limited performance. In this paper, we present a simple, but effective solution for semantic correspondence, called SemiMatch, that learns the networks in a semi-supervised manner by supplementing few ground-truth correspondences via utilization of a large amount of confident correspondences as pseudo-labels. Specifically, our framework generates the pseudo-labels using the model's prediction itself between source and weakly-augmented target, and uses pseudo-labels to learn the model again between source and strongly-augmented target, which improves the robustness of the model. We also present a novel confidence measure for pseudo-labels and data augmentation tailored for semantic correspondence. In experiments, SemiMatch achieves state-of-the-art performance on various benchmarks by a large margin.
****
Learned Queries for Efficient Local Attention
Moab Arar, Ariel Shamir, Amit H. Bermano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10841-10852
Vision Transformers (ViT) serve as powerful vision models. Unlike convolutional neural networks, which dominated vision research in previous years, vision transformers enjoy the ability to capture long-range dependencies in the data. Nonetheless, an integral part of any transformer architecture, the self-attention mechanism, suffers from high latency and inefficient memory utilization, making it less suitable for high-resolution input images. To alleviate these shortcomings, hierarchical vision models locally employ self-attention on non-interleaving windows. This relaxation reduces the complexity to be linear in the input size; however, it limits the cross-window interaction, hurting the model performance. In this paper, we propose a new shift-invariant local attention layer, called query and attend (QnA), that aggregates the input locally in an overlapping manner, much like convolutions. The key idea behind QnA is to introduce learned queries, which allow fast and efficient implementation. We verify the effectiveness of our layer by incorporating it into a hierarchical vision transformer model. We show improvements in speed and memory complexity while achieving comparable accurac

y with state-of-the-art models. Finally, our layer scales especially well with w
indow size, requiring up to x10 less memory while being up to x5 faster than exi
sting methods. The code is publicly available at https://github.com/moabarar/qna
.
****

Stereoscopic Universal Perturbations Across Different Architectures and Datasets
Zachary Berger, Parth Agrawal, Tian Yu Liu, Stefano Soatto, Alex Wong; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2022, pp. 15180-15190
We study the effect of adversarial perturbations of images on deep stereo matchi
ng networks for the disparity estimation task. We present a method to craft a si
ngle set of perturbations that, when added to any stereo image pair in a dataset
, can fool a stereo network to significantly alter the perceived scene geometry.
 Our perturbation images are "universal" in that they not only corrupt estimates
 of the network on the dataset they are optimized for, but also generalize to di
fferent architectures trained on different datasets. We evaluate our approach on
 multiple benchmark datasets where our perturbations can increase the D1-error (
akin to fooling rate) of state-of-the-art stereo networks from 1% to as much as
87%. We investigate the effect of perturbations on the estimated scene geometry
and identify object classes that are most vulnerable. Our analysis on the activa
tions of registered points between left and right images led us to find architec
tural components that can increase robustness against adversaries. By simply des
igning networks with such components, one can reduce the effect of adversaries b
y up to 60.5%, which rivals the robustness of networks fine-tuned with costly ad
versarial data augmentation. Our design principle also improves their robustness
 against common image corruptions by an average of 70%.
****

Colar: Effective and Efficient Online Action Detection by Consulting Exemplars
Le Yang, Junwei Han, Dingwen Zhang; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2022, pp. 3160-3169
Online action detection has attracted increasing research interests in recent ye
ars. Current works model historical dependencies and anticipate the future to pe
rceive the action evolution within a video segment and improve the detection acc
uracy. However, the existing paradigm ignores category-level modeling and does n
ot pay sufficient attention to efficiency. Considering a category, its represent
ative frames exhibit various characteristics. Thus, the category-level modeling
can provide complimentary guidance to the temporal dependencies modeling. This p
aper develops an effective exemplar-consultation mechanism that first measures t
he similarity between a frame and exemplary frames, and then aggregates exemplar
y features based on the similarity weights. This is also an efficient mechanism,
 as both similarity measurement and feature aggregation require limited computat
ions. Based on the exemplar-consultation mechanism, the long-term dependencies c
an be captured by regarding historical frames as exemplars, while the category-l
evel modeling can be achieved by regarding representative frames from a category
 as exemplars. Due to the complementarity from the category-level modeling, our
method employs a lightweight architecture but achieves new high performance on t
hree benchmarks. In addition, using a spatio-temporal network to tackle video fr
ames, our method makes a good trade-off between effectiveness and efficiency. Co
de is available at https://github.com/VividLe/Online-Action-Detection.
****

AutoGPart: Intermediate Supervision Search for Generalizable 3D Part Segmentatio
n
Xueyi Liu, Xiaomeng Xu, Anyi Rao, Chuang Gan, Li Yi; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11624-1
1634
Training a generalizable 3D part segmentation network is quite challenging but o
f great importance in real-world applications. To tackle this problem, some work
s design task-specific solutions by translating human understanding of the task
to machine's learning process, which faces the risk of missing the optimal strat
egy since machines do not necessarily understand in the exact human way. Others

try to use conventional task-agnostic approaches designed for domain generalizat ion problems with no task prior knowledge considered. To solve the above issues, we propose AutoGPart, a generic method enabling training generalizable 3D part segmentation networks with the task prior considered. AutoGPart builds a supervi sion space with geometric prior knowledge encoded, and lets the machine to searc h for the optimal supervisions from the space for a specific segmentation task a utomatically. Extensive experiments on three generalizable 3D part segmentation tasks are conducted to demonstrate the effectiveness and versatility of AutoGPar t. We demonstrate that the performance of segmentation networks using simple bac kbones can be significantly improved when trained with supervisions searched by our method.

****

DeltaCNN: End-to-End CNN Inference of Sparse Frame Differences in Videos

Mathias Parger, Chengcheng Tang, Christopher D. Twigg, Cem Keskin, Robert Wang, Markus Steinberger; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2022, pp. 12497-12506

Convolutional neural network inference on video data requires powerful hardware for real-time processing. Given the inherent coherence across consecutive frames , large parts of a video typically change little. By skipping identical image re gions and truncating insignificant pixel updates, computational redundancy can i n theory be reduced significantly. However, these theoretical savings have been difficult to translate into practice, as sparse updates hamper computational con sistency and memory access coherence; which are key for efficiency on real hardw are. With DeltaCNN, we present a sparse convolutional neural network framework t hat enables sparse frame-by-frame updates to accelerate video inference in pract ice. We provide sparse implementations for all typical CNN layers and propagate sparse feature updates end-to-end - without accumulating errors over time. Delta CNN is applicable to all convolutional neural networks without retraining. To th e best of our knowledge, we are the first to significantly outperform the dense reference, cuDNN, in practical settings, achieving speedups of up to 7x with onl y marginal differences in accuracy.

****

HLRTF: Hierarchical Low-Rank Tensor Factorization for Inverse Problems in Multi-Dimensional Imaging

Yisi Luo, Xi-Le Zhao, Deyu Meng, Tai-Xiang Jiang; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19303-1931 2

Inverse problems in multi-dimensional imaging, e.g., completion, denoising, and compressive sensing, are challenging owing to the big volume of the data and the inherent ill-posedness. To tackle these issues, this work unsupervisedly learns a hierarchical low-rank tensor factorization (HLRTF) by solely using an observe d multi-dimensional image. Specifically, we embed a deep neural network (DNN) in to the tensor singular value decomposition framework and develop the HLRTF, whic h captures the underlying low-rank structures of multi-dimensional images with c ompact representation abilities. This DNN herein serves as a nonlinear transform from a vector to another to help obtain a better low-rank representation. Our H LRTF infers the parameters of the DNN and the underlying low-rank structure of t he original data from its observation via the gradient descent using a non-refer ence loss function in an unsupervised manner. To address the vanishing gradient in extreme scenarios, e.g., structural missing pixels, we introduce a parametric total variation regularization to constrain the DNN parameters and the tensor f actor parameters with theoretical analysis. We apply our HLRTF for typical inver se problems in multi-dimensional imaging including completion, denoising, and sn apshot spectral imaging, which demonstrates its generality and wide applicabilit y. Extensive results illustrate the superiority of our method as compared with s tate-of-the-art methods.

****

Leveraging Self-Supervision for Cross-Domain Crowd Counting

Weizhe Liu, Nikita Durasov, Pascal Fua; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5341-5352

State-of-the-art methods for counting people in crowded scenes rely on deep networks to estimate crowd density. While effective, these data-driven approaches rely on large amount of data annotation to achieve good performance, which stops these models from being deployed in emergencies during which data annotation is either too costly or cannot be obtained fast enough. One popular solution is to use synthetic data for training. Unfortunately, due to domain shift, the resulting models generalize poorly on real imagery. We remedy this shortcoming by training with both synthetic images, along with their associated labels, and unlabeled real images. To this end, we force our network to learn perspective-aware features by training it to recognize upside-down real images from regular ones and incorporate into it the ability to predict its own uncertainty so that it can generate useful pseudo labels for fine-tuning purposes. This yields an algorithm that consistently outperforms state-of-the-art cross-domain crowd counting ones without any extra computation at inference time.

****

MNSRNet: Multimodal Transformer Network for 3D Surface Super-Resolution

Wuyuan Xie, Tengcong Huang, Miaohui Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12703-12712

With the rapid development of display technology, it has become an urgent need to obtain realistic 3D surfaces with as high-quality as possible. Due to the unstructured and irregular nature of 3D object data, it is usually difficult to obtain high-quality surface details and geometry textures at a low cost. In this article, we propose an effective multimodal-driven deep neural network to perform 3D surface super-resolution in 2D normal domain, which is simple, accurate, and robust to the above difficulty. To leverage the multimodal information from different perspectives, we jointly consider the texture, depth, and normal modalities to simultaneously restore fine-grained surface details as well as preserve geometry structures. To better utilize the cross-modality information, we explore a two-bridge normal method with a transformer structure for feature alignment, and investigate an affine transform module for fusing multimodal features. Extensive experimental results on public and our newly constructed photometric stereo dataset demonstrate that the proposed method delivers promising surface geometry details compared with nine competitive schemes.

****

Gaussian Process Modeling of Approximate Inference Errors for Variational Autoencoders

Minyoung Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 244-253

Variational autoencoder (VAE) is a very successful generative model whose key element is the so called amortized inference network, which can perform test time inference using a single feed forward pass. Unfortunately, this comes at the cost of degraded accuracy in posterior approximation, often underperforming the instance-wise variational optimization. Although the latest semi-amortized approaches mitigate the issue by performing a few variational optimization updates starting from the VAE's amortized inference output, they inherently suffer from computational overhead for inference at test time. In this paper, we address the problem in a completely different way by considering a random inference model, where we model the mean and variance functions of the variational posterior as random Gaussian processes (GP). The motivation is that the deviation of the VAE's amortized posterior distribution from the true posterior can be regarded as random noise, which allows us to view the approximation error as uncertainty in posterior approximation that can be dealt with in a principled GP manner. In particular, our model can quantify the difficulty in posterior approximation by a Gaussian variational density. Inference in our GP model is done by a single feed forward pass through the network, significantly faster than semi-amortized methods. We show that our approach attains higher test data likelihood than the state-of-the-arts on several benchmark datasets.

****

PlaneMVS: 3D Plane Reconstruction From Multi-View Stereo

Jiachen Liu, Pan Ji, Nitin Bansal, Changjiang Cai, Qingan Yan, Xiaolei Huang, Yi

Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8665-8675
We present a novel framework named PlaneMVS for 3D plane reconstruction from multiple input views with known camera poses. Most previous learning-based plane reconstruction methods reconstruct 3D planes from single images, which highly rely on single-view regression and suffer from depth scale ambiguity. In contrast, we reconstruct 3D planes with a multi-view-stereo (MVS) pipeline that takes advantage of multi-view geometry. We decouple plane reconstruction into a semantic plane detection branch and a plane MVS branch. The semantic plane detection branch is based on a single-view plane detection framework but with differences. The plane MVS branch adopts a set of slanted plane hypotheses to replace conventional depth hypotheses to perform plane sweeping strategy and finally learns pixel-level plane parameters and its planar depth map. We present how the two branches are learned in a balanced way, and propose a soft-pooling loss to associate the outputs of the two branches and make them benefit from each other. Extensive experiments on various indoor datasets show that PlaneMVS significantly outperforms state-of-the-art (SOTA) single-view plane reconstruction methods on both plane detection and 3D geometry metrics. Our method even outperforms a set of SOTA learning-based MVS methods thanks to the learned plane priors. To the best of our knowledge, this is the first work on 3D plane reconstruction within an end-to-end MVS framework.
****
Scene Graph Expansion for Semantics-Guided Image Outpainting
Chiao-An Yang, Cheng-Yo Tan, Wan-Cyuan Fan, Cheng-Fu Yang, Meng-Lin Wu, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15617-15626
In this paper, we address the task of semantics-guided image outpainting, which is to complete an image by generating semantically practical content. Different from most existing image outpainting works, we approach the above task by understanding and completing image semantics at the scene graph level. In particular, we propose a novel network of Scene Graph Transformer (SGT), which is designed to take node and edge features as inputs for modeling the associated structural information. To better understand and process graph-based inputs, our SGT uniquely performs feature attention at both node and edge levels. While the former views edges as relationship regularization, the latter observes the co-occurrence of nodes for guiding the attention process. We demonstrate that, given a partial input image with its layout and scene graph, our SGT can be applied for scene graph expansion and its conversion to a complete layout. Following state-of-the-art layout-to-image conversions works, the task of image outpainting can be completed with sufficient and practical semantics introduced. Extensive experiments are conducted on the datasets of MS-COCO and Visual Genome, which quantitatively and qualitatively confirm the effectiveness of our proposed SGT and outpainting frameworks.
****
SoftGroup for 3D Instance Segmentation on Point Clouds
Thang Vu, Kookhoi Kim, Tung M. Luu, Thanh Nguyen, Chang D. Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2708-2717
Existing state-of-the-art 3D instance segmentation methods perform semantic segmentation followed by grouping. The hard predictions are made when performing semantic segmentation such that each point is associated with a single class. However, the errors stemming from hard decision propagate into grouping that results in (1) low overlaps between the predicted instance with the ground truth and (2) substantial false positives. To address the aforementioned problems, this paper proposes a 3D instance segmentation method referred to as SoftGroup by performing bottom-up soft grouping followed by top-down refinement. SoftGroup allows each point to be associated with multiple classes to mitigate the problems stemming from semantic prediction errors and suppresses false positive instances by learning to categorize them as background. Experimental results on different datasets and multiple evaluation metrics demonstrate the efficacy of SoftGroup. Its per

formance surpasses the strongest prior method by a significant margin of +6.2% o n the ScanNet v2 hidden test set and +6.8% on S3DIS Area 5 in terms of AP50. Sof tGroup is also fast, running at 345ms per scan with a single Titan X on ScanNet v2 dataset. The source code and trained models for both datasets are available a t https://github.com/thangvubk/SoftGroup.git.
****

SharpContour: A Contour-Based Boundary Refinement Approach for Efficient and Acc urate Instance Segmentation
Chenming Zhu, Xuanye Zhang, Yanran Li, Liangdong Qiu, Kai Han, Xiaoguang Han; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4392-4401
Excellent performance has been achieved on instance segmentation but the quality on the boundary area remains unsatisfactory, which leads to a rising attention on boundary refinement. For practical use, an ideal post-processing refinement s cheme are required to be accurate, generic and efficient. However, most of exist ing approaches propose pixel-wise refinement, which either introduce a massive c omputation cost or design specifically for different backbone models. Contour-ba sed models are efficient and generic to be incorporated with any existing segmen tation methods, but they often generate over-smoothed contour and tend to fail o n corner areas. In this paper, we propose an efficient contour-based boundary re finement approach, named SharpContour, to tackle the segmentation of boundary ar ea. We design a novel contour evolution process together with an Instance-aware Point Classifier. Our method deforms the contour iteratively by updating offsets in a discrete manner. Differing from existing contour evolution methods, SharpC ontour estimates each offset more independently so that it predicts much sharper and accurate contours. Notably, our method is generic to seamlessly work with d iverse existing models with a small computational cost. Experiments show that Sh arpContour achieves competitive gains whilst preserving high efficiency.
****

MVS2D: Efficient Multi-View Stereo via Attention-Driven 2D Convolutions
Zhenpei Yang, Zhile Ren, Qi Shan, Qixing Huang; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8574-8584
Deep learning has made significant impacts on multi-view stereo systems. State-o f-the-art approaches typically involve building a cost volume, followed by multi ple 3D convolution operations to recover the input image's pixel-wise depth. Whi le such end-to-end learning of plane-sweeping stereo advances public benchmarks' accuracy, they are typically very slow to compute. We present MVS2D, a highly e fficient multi-view stereo algorithm that seamlessly integrates multi-view const raints into single-view networks via an attention mechanism. Since MVS2Donly bui lds on 2D convolutions, it is at least 2x faster than all the notable counterpar ts. Moreover, our algorithm produces precise depth estimations and 3D reconstruc tions, achieving state-of-the-art results on challenging benchmarks ScanNet, SUN 3D, RGBD, and the classical DTU dataset. our algorithm also out-performs all oth er algorithms in the setting of inexact camera poses. Our code is released at ht tps://github.com/zhenpeiyang/MVS2D
****

FIBA: Frequency-Injection Based Backdoor Attack in Medical Image Analysis
Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, Dacheng Tao; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20876-20885
In recent years, the security of AI systems has drawn increasing research attent ion, especially in the medical imaging realm. To develop a secure medical image analysis (MIA) system, it is a must to study possible backdoor attacks (BAs), wh ich can embed hidden malicious behaviors into the system. However, designing a u nified BA method that can be applied to various MIA systems is challenging due t o the diversity of imaging modalities (e.g., X-Ray, CT, and MRI) and analysis ta sks (e.g., classification, detection, and segmentation). Most existing BA method s are designed to attack natural image classification models, which apply spatia l triggers to training images and inevitably corrupt the semantics of poisoned p ixels, leading to the failures of attacking dense prediction models. To address

this issue, we propose a novel Frequency-Injection based Backdoor Attack method (FIBA) that is capable of delivering attacks in various MIA tasks. Specifically, FIBA leverages a trigger function in the frequency domain that can inject the low-frequency information of a trigger image into the poisoned image by linearly combining the spectral amplitude of both images. Since it preserves the semantics of the poisoned image pixels, FIBA can perform attacks on both classification and dense prediction models. Experiments on three benchmarks in MIA (i.e., ISIC-2019 for skin lesion classification, KiTS-19 for kidney tumor segmentation, and EAD-2019 for endoscopic artifact detection), validate the effectiveness of FIBA and its superiority over state-of-the-art methods in attacking MIA models as well as bypassing backdoor defense. The code will be released.

****

Beyond Semantic to Instance Segmentation: Weakly-Supervised Instance Segmentation via Semantic Knowledge Transfer and Self-Refinement

Beomyoung Kim, YoungJoon Yoo, Chae Eun Rhee, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4278-4287

Weakly-supervised instance segmentation (WSIS) has been considered as a more challenging task than weakly-supervised semantic segmentation (WSSS). Compared to WSSS, WSIS requires instance-wise localization, which is difficult to extract from image-level labels. To tackle the problem, most WSIS approaches use off-the-shelf proposal techniques that require pre-training with instance or object level labels, deviating the fundamental definition of the fully-image-level supervised setting. In this paper, we propose a novel approach including two innovative components. First, we propose a semantic knowledge transfer to obtain pseudo instance labels by transferring the knowledge of WSSS to WSIS while eliminating the need for the off-the-shelf proposals. Second, we propose a self-refinement method to refine the pseudo instance labels in a self-supervised scheme and to use the refined labels for training in an online manner. Here, we discover an erroneous phenomenon, semantic drift, that occurred by the missing instances in pseudo instance labels categorized as background class. This semantic drift occurs confusion between background and instance in training and consequently degrades the segmentation performance. We term this problem as semantic drift problem and show that our proposed self-refinement method eliminates the semantic drift problem. The extensive experiments on PASCAL VOC 2012 and MS COCO demonstrate the effectiveness of our approach, and we achieve a considerable performance without off-the-shelf proposal techniques. The code is available at https://github.com/clovaai/BESTIE.

****

Bridged Transformer for Vision and Point Cloud 3D Object Detection

Yikai Wang, TengQi Ye, Lele Cao, Wenbing Huang, Fuchun Sun, Fengxiang He, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12114-12123

3D object detection is a crucial research topic in computer vision, which usually uses 3D point clouds as input in conventional setups. Recently, there is a trend of leveraging multiple sources of input data, such as complementing the 3D point cloud with 2D images that often have richer color and fewer noises. However, due to the heterogeneous geometrics of the 2D and 3D representations, it prevents us from applying off-the-shelf neural networks to achieve multimodal fusion. To that end, we propose Bridged Transformer (BrT), an end-to-end architecture for 3D object detection. BrT is simple and effective, which learns to identify 3D and 2D object bounding boxes from both points and image patches. A key element of BrT lies in the utilization of object queries for bridging 3D and 2D spaces, which unifies different sources of data representations in Transformer. We adopt a form of feature aggregation realized by point-to-patch projections which further strengthen the interaction between images and points. Moreover, BrT works seamlessly for fusing the point cloud with multi-view images. We experimentally show that BrT surpasses state-of-the-art methods on SUN RGB-D and ScanNetV2 datasets.

****

Deep Constrained Least Squares for Blind Image Super-Resolution
Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17642-17652
In this paper, we tackle the problem of blind image super-resolution(SR) with a reformulated degradation model and two novel modules. Following the common practices of blind SR, our method proposes to improve both the kernel estimation as well as the kernel-based high-resolution image restoration. To be more specific, we first reformulate the degradation model such that the deblurring kernel estimation can be transferred into the low-resolution space. On top of this, we introduce a dynamic deep linear filter module. Instead of learning a fixed kernel for all images, it can adaptively generate deblurring kernel weights conditional on the input and yield a more robust kernel estimation. Subsequently, a deep constrained least square filtering module is applied to generate clean features based on the reformulation and estimated kernel. The deblurred feature and the low input image feature are then fed into a dual-path structured SR network and restore the final high-resolution result. To evaluate our method, we further conduct evaluations on several benchmarks, including Gaussian8 and DIV2KRK. Our experiments demonstrate that the proposed method achieves better accuracy and visual improvements against state-of-the-art methods. Codes and models are available at https://github.com/megvii-research/DCLS-SR.
****
EDTER: Edge Detection With Transformer
Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1402-1412
Convolutional neural networks have made significant progresses in edge detection by progressively exploring the context and semantic features. However, local details are gradually suppressed with the enlarging of receptive fields. Recently, vision transformer has shown excellent capability in capturing long-range dependencies. Inspired by this, we propose a novel transformer-based edge detector, Edge Detection TransformER (EDTER), to extract clear and crisp object boundaries and meaningful edges by exploiting the full image context information and detailed local cues simultaneously. EDTER works in two stages. In Stage I, a global transformer encoder is used to capture long-range global context on coarse-grained image patches. Then in Stage II, a local transformer encoder works on fine-grained patches to excavate the short-range local cues. Each transformer encoder is followed by an elaborately designed Bi-directional Multi-Level Aggregation decoder to achieve high-resolution features. Finally, the global context and local cues are combined by a Feature Fusion Module and fed into a decision head for edge prediction. Extensive experiments on BSDS500, NYUDv2, and Multicue demonstrate the superiority of EDTER in comparison with state-of-the-arts. The source code is available at https://github.com/MengyangPu/EDTER.
****
Fine-Tuning Global Model via Data-Free Knowledge Distillation for Non-IID Federated Learning
Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, Ling-Yu Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10174-10183
Federated Learning (FL) is an emerging distributed learning paradigm under privacy constraint. Data heterogeneity is one of the main challenges in FL, which results in slow convergence and degraded performance. Most existing approaches only tackle the heterogeneity challenge by restricting the local model update in client, ignoring the performance drop caused by direct global model aggregation. Instead, we propose a data-free knowledge distillation method to fine-tune the global model in the server (FedFTG), which relieves the issue of direct model aggregation. Concretely, FedFTG explores the input space of local models through a generator, and uses it to transfer the knowledge from local models to the global model. Besides, we propose a hard sample mining scheme to achieve effective knowledge distillation throughout the training. In addition, we develop customized la

bel sampling and class-level ensemble to derive maximum utilization of knowledge , which implicitly mitigates the distribution discrepancy across clients. Extens ive experiments show that our FedFTG significantly outperforms the state-of-the- art (SOTA) FL algorithms and can serve as a strong plugin for enhancing FedAvg, FedProx, FedDyn, and SCAFFOLD.
****

JIFF: Jointly-Aligned Implicit Face Function for High Quality Single View Clothe d Human Reconstruction

Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, Kwan-Yee K. Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 , pp. 2729-2739

This paper addresses the problem of single view 3D human reconstruction. Recent implicit function based methods have shown impressive results, but they fail to recover fine face details in their reconstructions. This largely degrades user e xperience in applications like 3D telepresence. In this paper, we focus on impro ving the quality of face in the reconstruction and propose a novel Jointly-align ed Implicit Face Function (JIFF) that combines the merits of the implicit functi on based approach and model based approach. We employ a 3D morphable face model as our shape prior and compute space-aligned 3D features that capture detailed f ace geometry information. Such space-aligned 3D features are combined with pixel -aligned 2D features to jointly predict an implicit face function for high quali ty face reconstruction. We further extend our pipeline and introduce a coarse-to -fine architecture to predict high quality texture for our detailed face model. Extensive evaluations have been carried out on public datasets and our proposed JIFF has demonstrated superior performance (both quantitatively and qualitativel y) over existing state-of-the-arts.
****

Deep 3D-to-2D Watermarking: Embedding Messages in 3D Meshes and Extracting Them From 2D Renderings

Innfarn Yoo, Huiwen Chang, Xiyang Luo, Ondrej Stava, Ce Liu, Peyman Milanfar, Fe ng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2022, pp. 10031-10040

Digital watermarking is widely used for copyright protection. Traditional 3D wat ermarking approaches or commercial software are typically designed to embed mess ages into 3D meshes, and later retrieve the messages directly from distorted/und istorted watermarked 3D meshes. However, in many cases, users only have access t o rendered 2D images instead of 3D meshes. Unfortunately, retrieving messages fr om 2D renderings of 3D meshes is still challenging and underexplored. We introdu ce a novel end-to-end learning framework to solve this problem through: 1) an en coder to covertly embed messages in both mesh geometry and textures; 2) a differ entiable renderer to render watermarked 3D objects from different camera angles and under varied lighting conditions; 3) a decoder to recover the messages from 2D rendered images. From our experiments, we show that our model can learn to em bed information visually imperceptible to humans, and to retrieve the embedded i nformation from 2D renderings that undergo 3D distortions. In addition, we demon strate that our method can also work with other renderers, such as ray tracers a nd real-time renderers with and without fine-tuning.
****

Beyond a Pre-Trained Object Detector: Cross-Modal Textual and Visual Context for Image Captioning

Chia-Wen Kuo, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2022, pp. 17969-17979

Significant progress has been made on visual captioning, largely relying on pre- trained features and later fixed object detectors that serve as rich inputs to a uto-regressive models. A key limitation of such methods, however, is that the ou tput of the model is conditioned only on the object detector's outputs. The assu mption that such outputs can represent all necessary information is unrealistic, especially when the detector is transferred across datasets. In this work, we r eason about the graphical model induced by this assumption, and propose to add a n auxiliary input to represent missing information such as object relationships.

We specifically propose to mine attributes and relationships from the Visual Genome dataset and condition the captioning model on them. Crucially, we propose (and show to be important) the use of a multi-modal pre-trained model (CLIP) to retrieve such contextual descriptions. Further, the object detector outputs are fixed due to a frozen model and hence do not have sufficient richness to allow the captioning model to properly ground them. As a result, we propose to condition both the detector and description outputs on the image, and show qualitatively that this can improve grounding. We validate our method on image captioning, perform thorough analyses of each component and importance of the pre-trained multi-modal model, and demonstrate significant improvements over the current state of the art, specifically +7.5% in CIDEr and +1.3% in BLEU-4 metrics.

****

Symmetry-Aware Neural Architecture for Embodied Visual Exploration

Shuang Liu, Takayuki Okatani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17242-17251

Visual exploration is a task that seeks to visit all the navigable areas of an environment as quickly as possible. The existing methods employ deep reinforcement learning (RL) as the standard tool for the task. However, they tend to be vulnerable to statistical shifts between the training and test data, resulting in poor generalization over novel environments that are out-of-distribution (OOD) from the training data. In this paper, we attempt to improve the generalization ability by utilizing the inductive biases available for the task. Employing the active neural SLAM (ANS) that learns exploration policies with the advantage actor-critic (A2C) method as the base framework, we first point out that the mappings represented by the actor and the critic should satisfy specific symmetries. We then propose a network design for the actor and the critic to inherently attain these symmetries. Specifically, we use G-convolution instead of the standard convolution and insert the semi-global polar pooling (SGPP) layer, which we newly design in this study, in the last section of the critic network. Experimental results show that our method increases area coverage by 8.1 square meters when trained on the Gibson dataset and tested on the Matterport3D dataset, establishing the new state-of-the-art.

****

AirObject: A Temporally Evolving Graph Embedding for Object Identification

Nikhil Varma Keetha, Chen Wang, Yuheng Qiu, Kuan Xu, Sebastian Scherer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8407-8416

Object encoding and identification are vital for robotic tasks such as autonomous exploration, semantic scene understanding, and re-localization. Previous approaches have attempted to either track objects or generate descriptors for object identification. However, such systems are limited to a "fixed" partial object representation from a single viewpoint. In a robot exploration setup, there is a requirement for a temporally "evolving" global object representation built as the robot observes the object from multiple viewpoints. Furthermore, given the vast distribution of unknown novel objects in the real world, the object identification process must be class-agnostic. In this context, we propose a novel temporal 3D object encoding approach, dubbed AirObject, to obtain global keypoint graph-based embeddings of objects. Specifically, the global 3D object embeddings are generated using a temporal convolutional network across structural information of multiple frames obtained from a graph attention-based encoding method. We demonstrate that AirObject achieves the state-of-the-art performance for video object identification and is robust to severe occlusion, perceptual aliasing, viewpoint shift, deformation, and scale transform, outperforming the state-of-the-art single-frame and sequential descriptors. To the best of our knowledge, AirObject is one of the first temporal object encoding methods. Source code is available at https://github.com/Nik-V9/AirObject.

****

From Representation to Reasoning: Towards Both Evidence and Commonsense Reasoning for Video Question-Answering

Jiangtong Li, Li Niu, Liqing Zhang; Proceedings of the IEEE/CVF Conference on Co

mputer Vision and Pattern Recognition (CVPR), 2022, pp. 21273-21282
Video understanding has achieved great success in representation learning, such as video caption, video object grounding, and video descriptive question-answer. However, current methods still struggle on video reasoning, including evidence reasoning and commonsense reasoning. To facilitate deeper video understanding towards video reasoning, we present the task of Causal-VidQA, which includes four types of questions ranging from scene description (description) to evidence reasoning (explanation) and commonsense reasoning (prediction and counterfactual). For commonsense reasoning, we set up a two-step solution by answering the question and providing a proper reason. Through extensive experiments on existing VideoQA methods, we find that the state-of-the-art methods are strong in descriptions but weak in reasoning. We hope that Causal-VidQA can guide the research of video understanding from representation learning to deeper reasoning. The dataset and related resources are available at https://github.com/bcmi/Causal-VidQA.git.
****

Semantic-Aware Domain Generalized Segmentation
Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, Wen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2594-2605
Deep models trained on source domain lack generalization when evaluated on unseen target domains with different data distributions. The problem becomes even more pronounced when we have no access to target domain samples for adaptation. In this paper, we address domain generalized semantic segmentation, where a segmentation model is trained to be domain-invariant without using any target domain data. Existing approaches to tackle this problem standardize data into a unified distribution. We argue that while such a standardization promotes global normalization, the resulting features are not discriminative enough to get clear segmentation boundaries. To enhance separation between categories while simultaneously promoting domain invariance, we propose a framework including two novel modules: Semantic-Aware Normalization (SAN) and Semantic-Aware Whitening (SAW). Specifically, SAN focuses on category-level center alignment between features from different image styles, while SAW enforces distributed alignment for the already center-aligned features. With the help of SAN and SAW, we encourage both intraclass compactness and inter-class separability. We validate our approach through extensive experiments on widely-used datasets (i.e. GTAV, SYNTHIA, Cityscapes, Mapillary and BDDS). Our approach shows significant improvements over existing state-of-the-art on various backbone networks. Code is available at https://github.com/leolyj/SAN-SAW
****

TransVPR: Transformer-Based Place Recognition With Multi-Level Attention Aggregation
Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13648-13657
Visual place recognition is a challenging task for applications such as autonomous driving navigation and mobile robot localization. Distracting elements presenting in complex scenes often lead to deviations in the perception of visual place. To address this problem, it is crucial to integrate information from only task-relevant regions into image representations. In this paper, we introduce a novel holistic place recognition model, TransVPR, based on vision Transformers. It benefits from the desirable property of the self-attention operation in Transformers which can naturally aggregate task-relevant features. Attentions from multiple levels of the Transformer, which focus on different regions of interest, are further combined to generate a global image representation. In addition, the output tokens from Transformer layers filtered by the fused attention mask are considered as key-patch descriptors, which are used to perform spatial matching to re-rank the candidates retrieved by the global image features. The whole model allows end-to-end training with a single objective and image-level supervision. TransVPR achieves state-of-the-art performance on several real-world benchmarks while maintaining low computational time and storage requirements.

****
DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion
Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20993-21002
A typical pipeline for multi-object tracking (MOT) is to use a detector for object localization, and following re-identification (re-ID) for object association. This pipeline is partially motivated by recent progress in both object detection and re-ID, and partially motivated by biases in existing tracking datasets, where most objects tend to have distinguishing appearance and re-ID models are sufficient for establishing associations. In response to such bias, we would like to re-emphasize that methods for multi-object tracking should also work when object appearance is not sufficiently discriminative. To this end, we propose a large-scale dataset for multi-human tracking, where humans have similar appearance, diverse motion and extreme articulation. As the dataset contains mostly group dancing videos, we name it "DanceTrack". We expect DanceTrack to provide a better platform to develop more MOT algorithms that rely less on visual discrimination and depend more on motion analysis. We benchmark several state-of-the-art trackers on our dataset and observe a significant performance drop on DanceTrack when compared against existing benchmarks. The dataset, project code and competition is released at: https://github.com/DanceTrack.
****
Unsupervised Learning of Debiased Representations With Pseudo-Attributes
Seonguk Seo, Joon-Young Lee, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16742-16751
The distributional shift issue between training and test sets is a critical challenge in machine learning, and is aggravated when models capture unintended decision rules with spurious correlations. Although existing works often handle this issue using human supervision, the availability of the proper annotations is impractical and even unrealistic. To better tackle this challenge, we propose a simple but effective debiasing technique in an unsupervised manner. Specifically, we perform clustering on the feature embedding space and identify pseudo-bias-attributes by taking advantage of the clustering results even without an explicit attribute supervision. Then, we employ a novel cluster-based reweighting scheme for learning debiased representation; this prevents minority groups from being ignored for minimizing the overall loss, which is desirable for worst-case generalization. The extensive experiments demonstrate the outstanding performance of our approach on multiple standard benchmarks, which is even as competitive as the supervised method. We plan to release the source code of our work for better reproducibility.
****
Protecting Celebrities From DeepFake With Identity Consistency Transformer
Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9468-9478
In this work we propose Identity Consistency Transformer, a novel face forgery detection method that focuses on high-level semantics, specifically identity information, and detecting a suspect face by finding identity inconsistency in inner and outer face regions. The Identity Consistency Transformer incorporates a consistency loss for identity consistency determination. We show that Identity Consistency Transformer exhibits superior generalization ability not only across different datasets but also across various types of image degradation forms found in real-world applications including deepfake videos. The Identity Consistency Transformer can be easily enhanced with additional identity information when such information is available, and for this reason it is especially well-suited for detecting face forgeries involving celebrities.
****
Give Me Your Attention: Dot-Product Attention Considered Harmful for Adversarial Patch Robustness
Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, Jan H

endrik Metzen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15234-15243
Neural architectures based on attention such as vision transformers are revolutionizing image recognition. Their main benefit is that attention allows reasoning about all parts of a scene jointly. In this paper, we show how the global reasoning of (scaled) dot-product attention can be the source of a major vulnerability when confronted with adversarial patch attacks. We provide a theoretical understanding of this vulnerability and relate it to an adversary's ability to misdirect the attention of all queries to a single key token under the control of the adversarial patch. We propose novel adversarial objectives for crafting adversarial patches which target this vulnerability explicitly. We show the effectiveness of the proposed patch attacks on popular image classification (ViTs and DeiTs) and object detection models (DETR). We find that adversarial patches occupying 0.5% of the input can lead to robust accuracies as low as 0% for ViT on ImageNet, and reduce the mAP of DETR on MS COCO to less than 3%.
****

TubeDETR: Spatio-Temporal Video Grounding With Transformers
Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16442-16453
We consider the problem of localizing a spatio-temporal tube in a video corresponding to a given text query. This is a challenging task that requires the joint and efficient modeling of temporal, spatial and multi-modal interactions. To address this task, we propose TubeDETR, a transformer-based architecture inspired by the recent success of such models for text-conditioned object detection. Our model notably includes: (i) an efficient video and text encoder that models spatial multi-modal interactions over sparsely sampled frames and (ii) a space-time decoder that jointly performs spatio-temporal localization. We demonstrate the advantage of our proposed components through an extensive ablation study. We also evaluate our full approach on the spatio-temporal video grounding task and demonstrate improvements over the state of the art on the challenging VidSTG and HC-STVG benchmarks.
****

KG-SP: Knowledge Guided Simple Primitives for Open World Compositional Zero-Shot Learning
Shyamgopal Karthik, Massimiliano Mancini, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9336-9345
The goal of open-world compositional zero-shot learning(OW-CZSL) is to recognize compositions of state and objects in images, given only a subset of them during training and no prior on the unseen compositions. In this setting, models operate on a huge output space, containing all possible state-object compositions. While previous works tackle the problem by learning embeddings for the compositions jointly,here we revisit a simple CZSL baseline and predict the primitives, i.e. states and objects, independently. To ensure that the model develops primitive-specific features, we equip the state and object classifiers with separate, non-linear feature extractors. Moreover, we estimate the feasibility of each composition through external knowledge, using this prior to remove unfeasible compositions from the output space.Finally, we propose a new setting, i.e. CZSL under partial supervision (pCZSL), where either only objects or state labels are available during training and we can use our prior to estimate the missing labels. Our model, Knowledge-Guided Simple Primitives (KG-SP), achieves the state of the art in both OW-CZSL and pCZSL, surpassing most recent competitors even when coupled with semi-supervised learning techniques
****

SLIC: Self-Supervised Learning With Iterative Clustering for Human Action Videos
Salar Hosseini Khorasgani, Yuxuan Chen, Florian Shkurti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16091-16101
Self-supervised methods have significantly closed the gap with end-to-end superv

ised learning for image classification [13,24]. In the case of human action vide os, however, where both appearance and motion are significant factors of variati on, this gap remains significant [28,58]. One of the key reasons for this is tha t sampling pairs of similar video clips, a required step for many self-supervise d contrastive learning methods, is currently done conservatively to avoid false positives. A typical assumption is that similar clips only occur temporally clos e within a single video, leading to insufficient examples of motion similarity. To mitigate this, we propose SLIC, a clustering-based self-supervised contrastiv e learning method for human action videos. Our key contribution is that we impro ve upon the traditional intra-video positive sampling by using iterative cluster ing to group similar video instances. This enables our method to leverage pseudo -labels from the cluster assignments to sample harder positives and negatives. S LIC outperforms state-of-the-art video retrieval baselines by +15.4% on top-1 re call on UCF101 and by +5.7% when directly transferred to HMDB51. With end-to-end finetuning for action classification, SLIC achieves 83.2% top-1 accuracy (+0.8%) on UCF101 and 54.5% on HMDB51 (+1.6%). SLIC is also competitive with the state -of-the-art in action classification after self-supervised pretraining on Kineti cs400.

****

CD2-pFed: Cyclic Distillation-Guided Channel Decoupling for Model Personalizatio n in Federated Learning

Yiqing Shen, Yuyin Zhou, Lequan Yu; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2022, pp. 10041-10050

Federated learning (FL) is a distributed learning paradigm that enables multiple clients to collaboratively learn a shared global model. Despite the recent prog ress, it remains challenging to deal with heterogeneous data clients, as the dis crepant data distributions usually prevent the global model from delivering good generalization ability on each participating client. In this paper, we propose CD^2-pFed, a novel Cyclic Distillation-guided Channel Decoupling framework, to p ersonalize the global model in FL, under various settings of data heterogeneity. Different from previous works which establish layer-wise personalization to ove rcome the non-IID data across different clients, we make the first attempt at ch annel-wise assignment for model personalization, referred to as channel decoupli ng. To further facilitate the collaboration between private and shared weights, we propose a novel cyclic distillation scheme to impose a consistent regularizat ion between the local and global model representations during the federation. Gu ided by the cyclical distillation, our channel decoupling framework can deliver more accurate and generalized results for different kinds of heterogeneity, such as feature skew, label distribution skew, and concept shift. Comprehensive expe riments on four benchmarks, including natural image and medical image analysis t asks, demonstrate the consistent effectiveness of our method on both local and e xternal validations.

****

UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Su medrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp . 20143-20153

Detecting abnormal events in video is commonly framed as a one-class classificat ion task, where training videos contain only normal events, while test videos en compass both normal and abnormal events. In this scenario, anomaly detection is an open-set problem. However, some studies assimilate anomaly detection to actio n recognition. This is a closed-set scenario that fails to test the capability o f systems at detecting new anomaly types. To this end, we propose UBnormal, a ne w supervised open-set benchmark composed of multiple virtual scenes for video an omaly detection. Unlike existing data sets, we introduce abnormal events annotat ed at the pixel level at training time, for the first time enabling the use of f ully-supervised learning methods for abnormal event detection. To preserve the t ypical open-set formulation, we make sure to include disjoint sets of anomaly ty pes in our training and test collections of videos. To our knowledge, UBnormal i

s the first video anomaly detection benchmark to allow a fair head-to-head compa
rison between one-class open-set models and supervised closed-set models, as sho
wn in our experiments. Moreover, we provide empirical evidence showing that UBno
rmal can enhance the performance of a state-of-the-art anomaly detection framewo
rk on two prominent data sets, Avenue and ShanghaiTech. Our benchmark is freely
available at https://github.com/lilygeorgescu/UBnormal.
****

Beyond Cross-View Image Retrieval: Highly Accurate Vehicle Localization Using Sa
tellite Image
Yujiao Shi, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2022, pp. 17010-17020
This paper addresses the problem of vehicle-mounted camera localization by match
ing a ground-level image with an overhead-view satellite map. Existing methods o
ften treat this problem as cross-view image retrieval, and use learned deep feat
ures to match the ground-level query image to a partition (e.g., a small patch)
of the satellite map. By these methods, the localization accuracy is limited by
the partitioning density of the satellite map (often in the order of tens meters
). Departing from the conventional wisdom of image retrieval, this paper present
s a novel solution that can achieve highly-accurate localization. The key idea i
s to formulate the task as pose estimation and solve it by neural-net based opti
mization. Specifically, we design a two-branch CNN to extract robust features fr
om the ground and satellite images, respectively. To bridge the vast cross-view
domain gap, we resort to a Geometry Projection module that projects features fro
m the satellite map to the ground-view, based on a relative camera pose. Aiming
to minimize the differences between the projected features and the observed feat
ures, we employ a differentiable Levenberg-Marquardt (LM) module to search for t
he optimal camera pose iteratively. The entire pipeline is differentiable and ru
ns end-to-end. Extensive experiments on standard autonomous vehicle localization
 datasets have confirmed the superiority of the proposed method. Notably, e.g.,
starting from a coarse estimate of camera location within a wide region of 40m x
 40m, with an 80% likelihood our method quickly reduces the lateral location err
or to be within 5m on a new KITTI cross-view dataset.
****

Closing the Generalization Gap of Cross-Silo Federated Medical Image Segmentatio
n
An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R. Roth, Ali Hatamizadeh, Can Zh
ao, Daguang Xu, Heng Huang, Ziyue Xu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20866-20875
Cross-silo federated learning (FL) has attracted much attention in medical imagi
ng analysis with deep learning in recent years as it can resolve the critical is
sues of insufficient data, data privacy, and training efficiency. However, there
 can be a generalization gap between the model trained from FL and the one from
centralized training. This important issue comes from the non-iid data distribut
ion of the local data in the participating clients and is well-known as client d
rift. In this work, we propose a novel training framework FedSM to avoid the cli
ent drift issue and successfully close the generalization gap compared with the
centralized training for medical image segmentation tasks for the first time. We
 also propose a novel personalized FL objective formulation and a new method Sof
tPull to solve it in our proposed framework FedSM. We conduct rigorous theoretic
al analysis to guarantee its convergence for optimizing the non-convex smooth ob
jective function. Real-world medical image segmentation experiments using deep F
L validate the motivations and effectiveness of our proposed method.
****

AKB-48: A Real-World Articulated Object Knowledge Base
Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, Cewu Lu; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2022, pp. 14809-14818
Human life is populated with articulated objects. A comprehensive understanding
of articulated objects, namely appearance, structure, physics property, and sema
ntics, will benefit many research communities. As current articulated object und

erstanding solutions are usually based on synthetic object dataset with CAD models without physics properties, which prevent satisfied generalization from simulation to real-world applications in visual and robotics tasks. To bridge the gap, we present AKB-48: a large-scale Articulated object Knowledge Base which consists of 2,037 real-world 3D articulated object models of 48 categories. Each object is described by a knowledge graph ArtiKG. To build the AKB-48, we present a fast articulation knowledge modeling (FArM) pipeline, which can fulfill the ArtiKG for an articulated object within 10-15 minutes, and largely reduce the cost for object modeling in the real world. Using our dataset, we propose AKBNet, an integral pipeline for Category-level Visual Articulation Manipulation (C-VAM) task, in which we benchmark three sub-tasks, namely pose estimation, object reconstruction and manipulation. Dataset, codes, and models are publicly available at https://liuliu66.github.io/AKB-48.

****

Style-ERD: Responsive and Coherent Online Motion Style Transfer

Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, Michiel van de Panne; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6593-6603

Motion style transfer is a common method for enriching character animation. Motion style transfer algorithms are often designed for offline settings where motions are processed in segments. However, for online animation applications, such as real-time avatar animation from motion capture, motions need to be processed as a stream with minimal latency. In this work, we realize a flexible, high-quality motion style transfer method for this setting. We propose a novel style transfer model, Style-ERD, to stylize motions in an online manner with an Encoder-Recurrent-Decoder structure, along with a novel discriminator that combines feature attention and temporal attention. Our method stylizes motions into multiple target styles with a unified model. Although our method targets online settings, it outperforms previous offline methods in motion realism and style expressiveness and provides significant gains in runtime efficiency.

****

Leverage Your Local and Global Representations: A New Self-Supervised Learning Strategy

Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16580-16589

Self-supervised learning (SSL) methods aim to learn view-invariant representations by maximizing the similarity between the features extracted from different crops of the same image regardless of cropping size and content. In essence, this strategy ignores the fact that two crops may truly contain different image information, e.g., background and small objects, and thus tends to restrain the diversity of the learned representations. In this work, we address this issue by introducing a new self-supervised learning strategy, LoGo, that explicitly reasons about Local and Global crops. To achieve view invariance, LoGo encourages similarity between global crops from the same image, as well as between a global and a local crop. However, to correctly encode the fact that the content of smaller crops may differ entirely, LoGo promotes two local crops to have dissimilar representations, while being close to global crops. Our LoGo strategy can easily be applied to existing SSL methods. Our extensive experiments on a variety of datasets and using different self-supervised learning frameworks validate its superiority over existing approaches. Noticeably, we achieve better results than supervised models on transfer learning when using only 1/10 of the data.

****

Stratified Transformer for 3D Point Cloud Segmentation

Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8500-8509

3D point cloud segmentation has made tremendous progress in recent years. Most current methods focus on aggregating local features, but fail to directly model long-range dependencies. In this paper, we propose Stratified Transformer that is

able to capture long-range contexts and demonstrates strong generalization abil
ity and high performance. Specifically, we first put forward a novel key samplin
g strategy. For each query point, we sample nearby points densely and distant po
ints sparsely as its keys in a stratified way, which enables the model to enlarg
e the effective receptive field and enjoy long-range contexts at a low computati
onal cost. Also, to combat the challenges posed by irregular point arrangements,
 we propose first-layer point embedding to aggregate local information, which fa
cilitates convergence and boosts performance. Besides, we adopt contextual relat
ive position encoding to adaptively capture position information. Finally, a mem
ory-efficient implementation is introduced to overcome the issue of varying poin
t numbers in each window. Extensive experiments demonstrate the effectiveness an
d superiority of our method on S3DIS, ScanNetv2 and ShapeNetPart datasets. Code
is available at https://github.com/dvlab-research/Stratified-Transformer.
****

NeRF in the Dark: High Dynamic Range View Synthesis From Noisy Raw Images
Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jona
than T. Barron; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2022, pp. 16190-16199
Neural Radiance Fields (NeRF) is a technique for high quality novel view synthes
is from a collection of posed input images. Like most view synthesis methods, Ne
RF uses tonemapped low dynamic range (LDR) as input; these images have been proc
essed by a lossy camera pipeline that smooths detail, clips highlights, and dist
orts the simple noise distribution of raw sensor data. We modify NeRF to instead
 train directly on linear raw images, preserving the scene's full dynamic range.
 By rendering raw output images from the resulting NeRF, we can perform novel hi
gh dynamic range (HDR) view synthesis tasks. In addition to changing the camera
viewpoint, we can manipulate focus, exposure, and tonemapping after the fact. Al
though a single raw image appears significantly more noisy than a postprocessed
one, we show that NeRF is highly robust to the zero-mean distribution of raw noi
se. When optimized over many noisy raw inputs (25-200), NeRF produces a scene re
presentation so accurate that its rendered novel views outperform dedicated sing
le and multi-image deep raw denoisers run on the same wide baseline input images
. As a result, our method, which we call RawNeRF, can reconstruct scenes from ex
tremely noisy images captured in near-darkness.
****

DArch: Dental Arch Prior-Assisted 3D Tooth Instance Segmentation With Weak Annot
ations
Liangdong Qiu, Chongjie Ye, Pei Chen, Yunbi Liu, Xiaoguang Han, Shuguang Cui; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2022, pp. 20752-20761
Automatic tooth instance segmentation on 3D dental models is a fundamental task
for computer-aided orthodontic treatments. Existing learning-based methods rely
heavily on expensive point-wise annotations. To alleviate this problem, we are t
he first to explore a low-cost annotation way for 3D tooth instance segmentation
, i.e., labeling all tooth centroids and only a few teeth for each dental model.
 Regarding the challenge when only weak annotation is provided, we present a den
tal arch prior-assisted 3D tooth segmentation method, namely DArch. Our DArch co
nsists of two stages, including tooth centroid detection and tooth instance segm
entation. Accurately detecting the tooth centroids can help locate the individua
l tooth, thus benefiting the segmentation. Thus, our DArch proposes to leverage
the dental arch prior to assist the detection. Specifically, we firstly propose
a coarse-to-fine method to estimate the dental arch, in which the dental arch is
 initially generated by Bezier curve regression and then a lightweight network i
s trained to refine it. With the estimated dental arch, we then propose a novel
Arch-aware Point Sampling (APS) method to assist the tooth centroid proposal gen
eration. Meantime, a segmentor is independently trained using a patch-based trai
ning strategy, aiming to segment a tooth instance from a 3D patch centered at th
e tooth centroid. Experimental results on 4,773 dental models have shown our DAr
ch can accurately segment each tooth of a dental model, and its performance is s
uperior to the state-of-the-art methods.

****
Task Decoupled Framework for Reference-Based Super-Resolution
Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, Dazhi He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5931-5940
Reference-based super-resolution(RefSR) has achieved impressive progress on the recovery of high-frequency details thanks to an additional reference high-resolution(HR) image input. Although the superiority compared with Single-Image Super-Resolution(SISR), existing RefSR methods easily result in the reference-underuse issue and the reference-misuse as shown in Fig.1. In this work, we deeply investigate the cause of the two issues and further propose a novel framework to mitigate them. Our studies find that the issues are mostly due to the improper coupled framework design of current methods. Those methods conduct the super-resolution task of the input low-resolution(LR) image and the texture transfer task from the reference image together in one module, easily introducing the interference between LR and reference features. Inspired by this finding, we propose a novel framework, which decouples the two tasks of RefSR, eliminating the interference between the LR image and the reference image. The super-resolution task upsamples the LR image leveraging only the LR image itself. The texture transfer task extracts and transfers abundant textures from the reference image to the coarsely upsampled result of the super-resolution task. Extensive experiments demonstrate clear improvements in both quantitative and qualitative evaluations over state-of-the-art methods.
****
Aug-NeRF: Training Stronger Neural Radiance Fields With Triple-Level Physically-Grounded Augmentations
Tianlong Chen, Peihao Wang, Zhiwen Fan, Zhangyang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15191-15202
Neural Radiance Field (NeRF) regresses a neural parameterized scene by differentially rendering multi-view images with ground-truth supervision. However, when interpolating novel views, NeRF often yields inconsistent and visually non-smooth geometric results, which we consider as a generalization gap between seen and unseen views. Recent advances in convolutional neural networks have demonstrated the promise of advanced robust data augmentations, either random or learned, in enhancing both in-distribution and out-of-distribution generalization. Inspired by that, we propose Augmented NeRF (Aug-NeRF), which for the first time brings the power of robust data augmentations into regularizing the NeRF training. Particularly, our proposal learns to seamlessly blend worst-case perturbations into three distinct levels of the NeRF pipeline with physical grounds, including (1) the input coordinates, to simulate imprecise camera parameters at image capture; (2) intermediate features, to smoothen the intrinsic feature manifold; and (3) pre-rendering output, to account for the potential degradation factors in the multi-view image supervision. Extensive results demonstrate that Aug-NeRF effectively boosts NeRF performance in both novel view synthesis (up to 1.5 dB PSNR gain) and underlying geometry reconstruction. Furthermore, thanks to the implicit smooth prior injected by the triple-level augmentations, Aug-NeRF can even recover scenes from heavily corrupted images, a highly challenging setting untackled before. Our codes are available in https://github.com/VITA-Group/Aug-NeRF.
****
RGB-Multispectral Matching: Dataset, Learning Methodology, Evaluation
Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, Luigi Di Stefano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15958-15968
We address the problem of registering synchronized color (RGB) and multi-spectral (MS) images featuring very different resolution by solving stereo matching correspondences. Purposely, we introduce a novel RGB-MS dataset framing 13 different scenes in indoor environments and providing a total of 34 image pairs annotated with semi-dense, high-resolution ground-truth labels in the form of disparity maps. To tackle the task, we propose a deep learning architecture trained in a s

elf-supervised manner by exploiting a further RGB camera, required only during t
raining data acquisition. In this setup, we can conveniently learn cross-modal m
atching in the absence of ground-truth labels by distilling knowledge from an ea
sier RGB-RGB matching task based on a collection of about 11K unlabeled image tr
iplets. Experiments show that the proposed pipeline sets a good performance bar
(1.16 pixels average registration error) for future research on this novel, chal
lenging task.
****
Id-Free Person Similarity Learning
Bing Shuai, Xinyu Li, Kaustav Kundu, Joseph Tighe; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14689-146
99
Learning a unified person detection and re-identification model is a key compone
nt of modern trackers. However, training such models usually relies on the avail
ability of training images / videos that are manually labeled with both person b
oxes and their identities. In this work, we explore training such a model by onl
y using person box annotations, thus removing the necessity of manually labeling
 a training dataset with additional person identity annotation as these are expe
nsive to collect. To this end, we present a contrastive learning framework to le
arn person similarity without using manually labeled identity annotations. First
, we apply image-level augmentation to images on public person detection dataset
s, based on which we learn a strong model for general person detection as well a
s for short-term person re-identification. To learn a model capable of longer-te
rm re-identification, we leverage the natural appearance evolution of each perso
n in videos to serve as instance-level appearance augmentation in our contrastiv
e loss formulation. Without access to the target dataset or person identity anno
tation, our model achieves competitive results compared to existing fully-superv
ised state-of-the-art methods on both person search and person tracking tasks. O
ur model also shows promising results for saving the annotation cost that is nee
ded to achieve a certain level of performance on the person search task.
****
Temporal Complementarity-Guided Reinforcement Learning for Image-to-Video Person
 Re-Identification
Wei Wu, Jiawei Liu, Kecheng Zheng, Qibin Sun, Zheng-Jun Zha; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp.
 7319-7328
Image-to-video person re-identification aims to retrieve the same pedestrian as
the image-based query from a video-based gallery set. Existing methods treat it
as a cross-modality retrieval task and learn the common latent embeddings from i
mage and video modalities, which are both less effective and efficient due to la
rge modality gap and redundant feature learning by utilizing all video frames. I
n this work, we first regard this task as point-to-set matching problem identica
l to human decision process, and propose a novel Temporal Complementarity-Guided
 Reinforcement Learning (TCRL) approach for image-to-video person re-identificat
ion. TCRL employs deep reinforcement learning to make sequential judgments on dy
namically selecting suitable amount of frames from gallery videos, and accumulat
e adequate temporal complementary information among these frames by the guidance
 of the query image, towards balancing efficiency and accuracy. Specifically, TC
RL formulates point-to-set matching procedure as Markov decision process, where
a sequential judgement agent measures the uncertainty between the query image an
d all historical frames at each time step, and verifies that sufficient compleme
ntary clues are accumulated for judgment (same or different) or one more frames
are requested to assist judgment. Moreover, TCRL maintains a sequential feature
extraction module with a complementary residual detector to dynamically suppress
 redundant salient regions and thoroughly mine diverse complementary clues among
 these selected frames for enhancing frame-level representation. Extensive exper
iments demonstrate the superiority of our method.
****
Globetrotter: Connecting Languages by Connecting Images
Dídac Surís, Dave Epstein, Carl Vondrick; Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16474-16484
Machine translation between many languages at once is highly challenging, since training with ground truth requires supervision between all language pairs, which is difficult to obtain. Our key insight is that, while languages may vary drastically, the underlying visual appearance of the world remains consistent. We introduce a method that uses visual observations to bridge the gap between languages, rather than relying on parallel corpora or topological properties of the representations. We train a model that aligns segments of text from different languages if and only if the images associated with them are similar and each image in turn is well-aligned with its textual description. We train our model from scratch on a new dataset of text in over fifty languages with accompanying images. Experiments show that our method outperforms previous work on unsupervised word and sentence translation using retrieval.
****

Fairness-Aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models
Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, Kui Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10379-10388
Prioritizing fairness is of central importance in artificial intelligence (AI) systems, especially for those societal applications, e.g., hiring systems should recommend applicants equally from different demographic groups, and risk assessment systems must eliminate racism in criminal justice. Existing efforts towards the ethical development of AI systems have leveraged data science to mitigate biases in the training set or introduced fairness principles into the training process. For a deployed AI system, however, it may not allow for retraining or tuning in practice. By contrast, we propose a more flexible approach, i.e., fairness-aware adversarial perturbation (FAAP), which learns to perturb input data to blind deployed models on fairness-related features, e.g., gender and ethnicity. The key advantage is that FAAP does not modify deployed models in terms of parameters and structures. To achieve this, we design a discriminator to distinguish fairness-related attributes based on latent representations from deployed models. Meanwhile, a perturbation generator is trained against the discriminator, such that no fairness-related features could be extracted from perturbed inputs. Exhaustive experimental evaluation demonstrates the effectiveness and superior performance of the proposed FAAP. In addition, FAAP is validated on real-world commercial deployments (inaccessible to model parameters), which shows the transferability of FAAP, foreseeing the potential of black-box adaptation.
****

Stochastic Backpropagation: A Memory Efficient Strategy for Training Video Models
Feng Cheng, Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Li, Wei Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8301-8310
We propose a memory efficient method, named Stochastic Backpropagation (SBP), for training deep neural networks on videos. It is based on the finding that gradients from incomplete execution for backpropagation can still effectively train the models with minimal accuracy loss, which attributes to the high redundancy of video. SBP keeps all forward paths but randomly and independently removes the backward paths for each network layer in each training step. It reduces the GPU memory cost by eliminating the need to cache activation values corresponding to the dropped backward paths, whose amount can be controlled by an adjustable keep-ratio. Experiments show that SBP can be applied to a wide range of models for video tasks, leading to up to 80.0% GPU memory saving and 10% training speedup with less than 1% accuracy drop on action recognition and temporal action detection.
****

Semantic-Shape Adaptive Feature Modulation for Semantic Image Synthesis
Zhengyao Lv, Xiaoming Li, Zhenxing Niu, Bing Cao, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022,

pp. 11214-11223
Recent years have witnessed substantial progress in semantic image synthesis, it is still challenging in synthesizing photo-realistic images with rich details. Most previous methods focus on exploiting the given semantic map, which just captures an object-level layout for an image. Obviously, a fine-grained part-level semantic layout will benefit object details generation, and it can be roughly inferred from an object's shape. In order to exploit the part-level layouts, we propose a Shape-aware Position Descriptor (SPD) to describe each pixel's positional feature, where object shape is explicitly encoded into the SPD feature. Furthermore, a Semantic-shape Adaptive Feature Modulation (SAFM) block is proposed to combine the given semantic map and our positional features to produce adaptively modulated features. Extensive experiments demonstrate that the proposed SPD and SAFM significantly improve the generation of objects with rich details. Moreover, our method performs favorably against the SOTA methods in terms of quantitative and qualitative evaluation. The source code and model are available at https://github.com/cszy98/SAFM.
****

Egocentric Scene Understanding via Multimodal Spatial Rectifier
Tien Do, Khiem Vuong, Hyun Soo Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2832-2841
In this paper, we study a problem of egocentric scene understanding, i.e., predicting depths and surface normals from an egocentric image. Egocentric scene understanding poses unprecedented challenges: (1) due to large head movements, the images are taken from non-canonical viewpoints (i.e., tilted images) where existing models of geometry prediction do not apply; (2) dynamic foreground objects including hands constitute a large proportion of visual scenes. These challenges limit the performance of the existing models learned from large indoor datasets, such as ScanNet and NYUv2, which comprise predominantly upright images of static scenes. We present a multimodal spatial rectifier that stabilizes the egocentric images to a set of reference directions, which allows learning a coherent visual representation. Unlike unimodal spatial rectifier that often produces excessive perspective warp for egocentric images, the multimodal spatial rectifier learns from multiple directions that can minimize the impact of the perspective warp. To learn visual representations of the dynamic foreground objects, we present a new dataset called EDINA (Egocentric Depth on everyday INdoor Activities) that comprises more than 500K synchronized RGBD frames and gravity directions. Equipped with the multimodal spatial rectifier and the EDINA dataset, our proposed method on single-view depth and surface normal estimation significantly outperforms the baselines not only on our EDINA dataset, but also on other popular egocentric datasets, such as First Person Hand Action (FPHA) and EPIC-KITCHENS.
****

Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels
Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, Xinyi Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4248-4257
The crux of semi-supervised semantic segmentation is to assign pseudo-labels to the pixels of unlabeled images. A common practice is to select the highly confident predictions as the pseudo ground-truth, but it leads to a problem that most pixels may be left unused due to their unreliability. We argue that every pixel matters to the model training. Intuitively, an unreliable prediction may get confused among the top classes (i.e., those with the highest probabilities), however, it should be confident about the pixel not belonging to the remaining classes. Hence, such a pixel can be convincingly treated as a negative sample to those most unlikely categories. Based on this insight, we develop an effective pipeline to make sufficient use of unlabeled data. We first separate reliable and unreliable pixels via the predicted entropy map, then push each unreliable pixel to a category-wise queue that consists of negative samples, and finally train the model with all candidate pixels. Considering the training evolution, where the prediction becomes more and more accurate, we adaptively adjust the threshold for the reliable-unreliable partition. Experimental results on various benchmarks and

training settings demonstrate the superiority of our approach over the state-of-the-art alternatives.
****

Day-to-Night Image Synthesis for Training Nighttime Neural ISPs
Abhijith Punnappurath, Abdullah Abuolaim, Abdelrahman Abdelhamed, Alex Levinshtein, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10769-10778
Many flagship smartphone cameras now use a dedicated neural image signal processor (ISP) to render noisy raw sensor images to the final processed output. Training nightmode ISP networks relies on large-scale datasets of image pairs with: (1) a noisy raw image captured with a short exposure and a high ISO gain; and (2) a ground truth low-noise raw image captured with a long exposure and low ISO that has been rendered through the ISP. Capturing such image pairs is tedious and time-consuming, requiring careful setup to ensure alignment between the image pairs. In addition, ground truth images are often prone to motion blur due to the long exposure. To address this problem, we propose a method that synthesizes nighttime images from daytime images. Daytime images are easy to capture, exhibit low-noise (even on smartphone cameras) and rarely suffer from motion blur. We outline a processing framework to convert daytime raw images to have the appearance of realistic nighttime raw images with different levels of noise. Our procedure allows us to easily produce aligned noisy and clean nighttime image pairs. We show the effectiveness of our synthesis framework by training neural ISPs for nightmode rendering. Furthermore, we demonstrate that using our synthetic nighttime images together with small amounts of real data (e.g., 5% to 10%) yields performance almost on par with training exclusively on real nighttime images. Our dataset and code are available at https://github.com/SamsungLabs/day-to-night.
****

Commonality in Natural Images Rescues GANs: Pretraining GANs With Generic and Privacy-Free Synthetic Data
Kyungjune Baek, Hyunjung Shim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7854-7864
Transfer learning for GANs successfully improves generation performance under low-shot regimes. However, existing studies show that the pretrained model using a single benchmark dataset is not generalized to various target datasets. More importantly, the pretrained model can be vulnerable to copyright or privacy risks as membership inference attack advances. To resolve both issues, we propose an effective and unbiased data synthesizer, namely Primitives-PS, inspired by the generic characteristics of natural images. Specifically, we utilize 1) the generic statistics on the frequency magnitude spectrum, 2) the elementary shape (i.e., image composition via elementary shapes) for representing the structure information, and 3) the existence of saliency as prior. Since our synthesizer only considers the generic properties of natural images, the single model pretrained on our dataset can be consistently transferred to various target datasets, and even outperforms the previous methods pretrained with the natural images in terms of Fr'echet inception distance. Extensive analysis, ablation study, and evaluations demonstrate that each component of our data synthesizer is effective and provide insights on the desirable nature of the pretrained model for the transferability of GANs.
****