Invertible Denoising Network: A Light Solution for Real Noise Removal
Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, Tom G
edeon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2021, pp. 13365-13374
Invertible networks have various benefits for image denoising since they are lig
htweight, information-lossless, and memory-saving during back-propagation. Howev
er, applying invertible models to remove noise is challenging because the input
is noisy, and the reversed output is clean, following two different distribution
s. We propose an invertible denoising network, InvDN, to address this challenge.
 InvDN transforms the noisy input into a low-resolution clean image and a latent
 representation containing noise. To discard noise and restore the clean image,
InvDN replaces the noisy latent representation with another one sampled from a p
rior distribution during reversion. The denoising performance of InvDN is better
 than all the existing competitive models, achieving a new state-of-the-art resu
lt for the SIDD dataset while enjoying less run time. Moreover, the size of InvD
N is far smaller, only having 4.2% of the number of parameters compared to the m
ost recently proposed DANet. Further, via manipulating the noisy latent represen
tation, InvDN is also able to generate noise more similar to the original one. O
ur code is available at: https://github.com/Yang-Liu1082/InvDN.git.
****
Greedy Hierarchical Variational Autoencoders for Large-Scale Video Prediction
Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, Chelsea Finn; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2021, pp. 2318-2328
A video prediction model that generalizes to diverse scenes would enable intelli
gent agents such as robots to perform a variety of tasks via planning with the m
odel. However, while existing video prediction models have produced promising re
sults on small datasets, they suffer from severe underfitting when trained on la
rge and diverse datasets. To address this underfitting challenge, we first obser
ve that the ability to train larger video prediction models is often bottlenecke
d by the memory constraints of GPUs or TPUs. In parallel, deep hierarchical late
nt variable models can produce higher quality predictions by capturing the multi
-level stochasticity of future observations, but end-to-end optimization of such
 models is notably difficult. Our key insight is that greedy and modular optimiz
ation of hierarchical autoencoders can simultaneously address both the memory co
nstraints and the optimization challenges of large-scale video prediction. We in
troduce Greedy Hierarchical Variational Autoencoders (GHVAEs), a method that lea
rns high-fidelity video predictions by greedily training each level of a hierarc
hical autoencoder. In comparison to state-of-the-art models, GHVAEs provide 17-5
5% gains in prediction performance on four video datasets, a 35-40% higher succe
ss rate on real robot tasks, and can improve performance monotonically by simply
 adding more modules.
****
Over-the-Air Adversarial Flickering Attacks Against Video Recognition Networks
Roi Pony, Itay Naeh, Shie Mannor; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2021, pp. 515-524
Deep neural networks for video classification, just like image classification ne
tworks, may be subjected to adversarial manipulation. The main difference betwee
n image classifiers and video classifiers is that the latter usually use tempora
l information contained within the video. In this work we present a manipulation
 scheme for fooling video classifiers by introducing a flickering temporal pertu
rbation that in some cases may be unnoticeable by human observers and is impleme
ntable in the real world. After demonstrating the manipulation of action classif
ication of single videos, we generalize the procedure to make universal adversar
ial perturbation, achieving high fooling ratio. In addition, we generalize the u
niversal perturbation and produce a temporal-invariant perturbation, which can b
e applied to the video without synchronizing the perturbation to the input. The
attack was implemented on several target models and the transferability of the a
ttack was demonstrated. These properties allow us to bridge the gap between simu
lated environment and real-world application, as will be demonstrated in this pa

per for the first time for an over-the-air flickering attack.
****
Encoder Fusion Network With Co-Attention Embedding for Referring Image Segmentat
ion
Guang Feng, Zhiwei Hu, Lihe Zhang, Huchuan Lu; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15506-15515
Recently, referring image segmentation has aroused widespread interest. Previous
 methods perform the multi-modal fusion between language and vision at the decod
ing side of the network. And, linguistic feature interacts with visual feature o
f each scale separately, which ignores the continuous guidance of language to mu
lti-scale visual features. In this work, we propose an encoder fusion network (E
FN), which transforms the visual encoder into a multi-modal feature learning net
work, and uses language to refine the multi-modal features progressively. Moreov
er, a co-attention mechanism is embedded in the EFN to realize the parallel upda
te of multi-modal features, which can promote the consistent of the cross-modal
information representation in the semantic space. Finally, we propose a boundary
 enhancement module (BEM) to make the network pay more attention to the fine str
ucture. The experiment results on four benchmark datasets demonstrate that the p
roposed approach achieves the state-of-the-art performance under different evalu
ation metrics without any post-processing.
****
Polka Lines: Learning Structured Illumination and Reconstruction for Active Ster
eo
Seung-Hwan Baek, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2021, pp. 5757-5767
Active stereo cameras that recover depth from structured light captures have bec
ome a cornerstone sensor modality for 3D scene reconstruction and understanding
tasks across application domains. Active stereo cameras project a pseudo-random
dot pattern on object surfaces to extract disparity independently of object text
ure. Such hand-crafted patterns are designed in isolation from the scene statist
ics, ambient illumination conditions, and the reconstruction method. In this wor
k, we propose a method to jointly learn structured illumination and reconstructi
on, parameterized by a diffractive optical element and a neural network, in an e
nd-to-end fashion. To this end, we introduce a differentiable image formation mo
del for active stereo, relying on both wave and geometric optics, and a trinocul
ar reconstruction network. The jointly optimized pattern, which we dub "Polka Li
nes," together with the reconstruction network, makes accurate active-stereo dep
th estimates across imaging conditions. We validate the proposed method in simul
ation and using with an experimental prototype, and we demonstrate several varia
nts of the Polka Lines patterns specialized to the illumination conditions.
****
Image Inpainting With External-Internal Learning and Monochromic Bottleneck
Tengfei Wang, Hao Ouyang, Qifeng Chen; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5120-5129
Although recent inpainting approaches have demonstrated significant improvement
with deep neural networks, they still suffer from artifacts such as blunt struct
ures and abrupt colors when filling in the missing regions. To address these iss
ues, we propose an external-internal inpainting scheme with a monochromic bottle
neck that helps image inpainting models remove these artifacts. In the external
learning stage, we reconstruct missing structures and details in the monochromic
 space to reduce the learning dimension. In the internal learning stage, we prop
ose a novel internal color propagation method with progressive learning strategi
es for consistent color restoration. Extensive experiments demonstrate that our
proposed scheme helps image inpainting models produce more structure-preserved a
nd visually compelling results.
****
Patch2Pix: Epipolar-Guided Pixel-Level Correspondences
Qunjie Zhou, Torsten Sattler, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4669-4678
The classical matching pipeline used for visual localization typically involves

three steps: (i) local feature detection and description, (ii) feature matching, and (iii) outlier rejection. Recently emerged correspondence networks propose to perform those steps inside a single network but suffer from low matching resolution due to the memory bottleneck. In this work, we propose a new perspective to estimate correspondences in a detect-to-refine manner, where we first predict patch-level match proposals and then refine them. We present Patch2Pix, a novel refinement network that refines match proposals by regressing pixel-level matches from the local regions defined by those proposals and jointly rejecting outlier matches with confidence scores. Patch2Pix is weakly supervised to learn correspondences that are consistent with the epipolar geometry of an input image pair. We show that our refinement network significantly improves the performance of correspondence networks on image matching, homography estimation, and localization tasks. In addition, we show that our learned refinement generalizes to fully-supervised methods without re-training, which leads us to state-of-the-art localization performance. The code is available at https://github.com/GrumpyZhou/patch2pix.
****
Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer
Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2898-2907
Occluded person re-identification (Re-ID) is a challenging task as persons are frequently occluded by various obstacles or other persons, especially in the crowd scenario. To address these issues, we propose a novel end-to-end Part-Aware Transformer (PAT) for occluded person Re-ID through diverse part discovery via a transformer encoder-decoder architecture, including a pixel context based transformer encoder and a part prototype based transformer decoder. The proposed PAT model enjoys several merits. First, to the best of our knowledge, this is the first work to exploit the transformer encoder-decoder architecture for occluded person Re-ID in a unified deep model. Second, to learn part prototypes well with only identity labels, we design two effective mechanisms including part diversity and part discriminability. Consequently, we can achieve diverse part discovery for occluded person Re-ID in a weakly supervised manner. Extensive experimental results on six challenging benchmarks for three tasks (occluded, partial and holistic Re-ID) demonstrate that our proposed PAT performs favorably against stat-of-the-art methods.
****
Counterfactual Zero-Shot and Open-Set Visual Recognition
Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15404-15414
We present a novel counterfactual framework for both Zero-Shot Learning (ZSL) and Open-Set Recognition (OSR), whose common challenge is generalizing to the unseen-classes by only training on the seen-classes. Our idea stems from the observation that the generated samples for unseen-classes are often out of the true distribution, which causes severe recognition rate imbalance between the seen-class (high) and unseen-class (low). We show that the key reason is that the generation is not Counterfactual Faithful, and thus we propose a faithful one, whose generation is from the sample-specific counterfactual question: What would the sample look like, if we set its class attribute to a certain class, while keeping its sample attribute unchanged? Thanks to the faithfulness, we can apply the Consistency Rule to perform unseen/seen binary classification, by asking: Would its counterfactual still look like itself? If "yes", the sample is from a certain class, and "no" otherwise. Through extensive experiments on ZSL and OSR, we demonstrate that our framework effectively mitigates the seen/unseen imbalance and hence significantly improves the overall performance. Note that this framework is orthogonal to existing methods, thus, it can serve as a new baseline to evaluate how ZSL/OSR models generalize. Codes are available at https://github.com/yue-zhongqi/gcm-cf.

****
Person30K: A Dual-Meta Generalization Network for Person Re-Identification
Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, Ling-Yu Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2123-2132
Recently, person re-identification (ReID) has vastly benefited from the surging waves of data-driven methods. However, these methods are still not reliable enough for real-world deployments, due to the insufficient generalization capability of the models learned on existing benchmarks that have limitations in multiple aspects, including limited data scale, capture condition variations, and appearance diversities. To this end, we collect a new dataset named Person30K with the following distinct features: 1) a very large scale containing 1.38 million images of 30K identities, 2) a large capture system containing 6,497 cameras deployed at 89 different sites, 3) abundant sample diversities including varied backgrounds and diverse person poses. Furthermore, we propose a domain generalization ReID method, dual-meta generalization network (DMG-Net), to exploit the merits of meta-learning in both the training procedure and the metric space learning. Concretely, we design a "learning then generalization evaluation" meta-training procedure and a meta-discrimination loss to enhance model generalization and discrimination capabilities. Comprehensive experiments validate the effectiveness of our DMG-Net. (Dataset and code will be released.)
****
Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition
Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, Tobias Fischer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14141-14152
Visual Place Recognition is a challenging task for robotics and autonomous systems, which must deal with the twin problems of appearance and viewpoint change in an always changing world. This paper introduces Patch-NetVLAD, which provides a novel formulation for combining the advantages of both local and global descriptor methods by deriving patch-level features from NetVLAD residuals. Unlike the fixed spatial neighborhood regime of existing local keypoint features, our method enables aggregation and matching of deep-learned local features defined over the feature-space grid. We further introduce a multi-scale fusion of patch features that have complementary scales (i.e. patch sizes) via an integral feature space and show that the fused features are highly invariant to both condition (season, structure, and illumination) and viewpoint (translation and rotation) changes. Patch-NetVLAD achieves state-of-the-art visual place recognition results in computationally limited scenarios, validated on a range of challenging real-world datasets, including winning the Facebook Mapillary Visual Place Recognition Challenge at ECCV2020. It is also adaptable to user requirements, with a speed-optimised version operating over an order of magnitude faster than the state-of-the-art. By combining superior performance with improved computational efficiency in a configurable framework, Patch-NetVLAD is well suited to enhance both stand-alone place recognition capabilities and the overall performance of SLAM systems.
****
Visually Informed Binaural Audio Generation without Binaural Audios
Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15485-15494
Stereophonic audio, especially binaural audio, plays an essential role in immersive viewing environments. Recent research has explored generating stereophonic audios guided by visual cues and multi-channel audio collections in a fully-supervised manner. However, due to the requirement of professional recording devices, existing datasets are limited in scale and variety, which impedes the generalization of supervised methods to real-world scenarios. In this work, we propose PseudoBinaural, an effective pipeline that is free of binaural recordings. The key insight is to carefully build pseudo visual-stereo pairs with mono data for training. Specifically, we leverage spherical harmonic decomposition and head-relat

ed impulse response (HRIR) to identify the relationship between the location of a sound source and the received binaural audio. Then in the visual modality, corresponding visual cues of the mono data are manually placed at sound source positions to form the pairs. Compared to fully-supervised paradigms, our binaural-recording-free pipeline shows great stability in the cross-dataset evaluation and comparable performance under subjective preference. Moreover, combined with binaural recorded data, our method is able to further boost the performance of binaural audio generation under supervised settings.
****
Dual Attention Guided Gaze Target Detection in the Wild
Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, Guangtao Zhai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11390-11399
Gaze target detection aims to infer where each person in a scene is looking. Existing works focus on 2D gaze and 2D saliency, but fail to exploit 3D contexts. In this work, we propose a three-stage method to simulate the human gaze inference behavior in 3D space. In the first stage, we introduce a coarse-to-fine strategy to robustly estimate a 3D gaze orientation from the head. The predicted gaze is decomposed into a planar gaze on the image plane and a depth-channel gaze. In the second stage, we develop a Dual Attention Module (DAM), which takes the planar gaze to produce the filed of view and masks interfering objects regulated by depth information according to the depth-channel gaze. In the third stage, we use the generated dual attention as guidance to perform two sub-tasks: (1) identifying whether the gaze target is inside or out of the image; (2) locating the target if inside. Extensive experiments demonstrate that our approach performs favorably against state-of-the-art methods on GazeFollow and VideoAttentionTarget datasets.
****
Privacy Preserving Localization and Mapping From Uncalibrated Cameras
Marcel Geppert, Viktor Larsson, Pablo Speciale, Johannes L. Schonberger, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1809-1819
Recent works on localization and mapping from privacy preserving line features have made significant progress towards addressing the privacy concerns arising from cloud-based solutions in mixed reality and robotics. The requirement for calibrated cameras is a fundamental limitation for these approaches, which prevents their application in many crowd-sourced mapping scenarios. In this paper, we propose a solution to the uncalibrated privacy preserving localization and mapping problem. Our approach simultaneously recovers the intrinsic and extrinsic calibration of a camera from line-features only. This enables uncalibrated devices to both localize themselves within an existing map as well as contribute to the map, while preserving the privacy of the image contents. Furthermore, we also derive a solution to bootstrapping maps from scratch using only uncalibrated devices. Our approach provides comparable performance to the calibrated scenario and the privacy compromising alternatives based on traditional point features.
****
Learning Calibrated Medical Image Segmentation via Multi-Rater Agreement Modeling
Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, Yefeng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12341-12351
In medical image analysis, it is typical to collect multiple annotations, each from a different clinical expert or rater, in the expectation that possible diagnostic errors could be mitigated. Meanwhile, from the computer vision practitioner viewpoint, it has been a common practice to adopt the ground-truth obtained via either the majority-vote or simply one annotation from a preferred rater. This process, however, tends to overlook the rich information of agreement or disagreement ingrained in the raw multi-rater annotations. To address this issue, we propose to explicitly model the multi-rater (dis-)agreement, dubbed MRNet, which has two main contributions. First, an expertise-aware inferring module or EIM is

devised to embed the expertise level of individual raters as prior knowledge, to form high-level semantic features. Second, our approach is capable of reconstructing multi-rater gradings from coarse predictions, with the multi-rater (dis-)agreement cues being further exploited to improve the segmentation performance. To our knowledge, our work is the first in producing calibrated predictions under different expertise levels for medical image segmentation. Extensive empirical experiments are conducted across five medical segmentation tasks of diverse imaging modalities. In these experiments, superior performance of our MRNet is observed comparing to the state-of-the-arts, indicating the effectiveness and applicability of our MRNet toward a wide range of medical segmentation tasks.
****
Points As Queries: Weakly Semi-Supervised Object Detection by Points
Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8823-8832
We propose a novel point annotated setting for the weakly semi-supervised object detection task, in which the dataset comprises small fully annotated images and large weakly annotated images by points. It achieves a balance between tremendous annotation burden and detection performance. Based on this setting, we analyze existing detectors and find that these detectors have difficulty in fully exploiting the power of the annotated points. To solve this, we introduce a new detector, Point DETR, which extends DETR by adding a point encoder. Extensive experiments conducted on MS-COCO dataset in various data settings show the effectiveness of our method. In particular, when using 20% fully labeled data from COCO, our detector achieves a promising performance, 33.3 AP, which outperforms a strong baseline (FCOS) by 2.0 AP, and we demonstrate the point annotations bring over 10 points in various AR metrics.
****
Removing Diffraction Image Artifacts in Under-Display Camera via Dynamic Skip Connection Network
Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, Jinwei Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 662-671
Recent development of Under-Display Camera (UDC) systems provides a true bezel-less and notch-free viewing experience on smartphones (and TV, laptops, tablets), while allowing images to be captured from the selfie camera embedded underneath. In a typical UDC system, the microstructure of the semi-transparent organic light-emitting diode (OLED) pixel array attenuates and diffracts the incident light on the camera, resulting in significant image quality degradation. Oftentimes, noise, flare, haze, and blur can be observed in UDC images. In this work, we aim to analyze and tackle the aforementioned degradation problems. We define a physics-based image formation model to better understand the degradation. In addition, we utilize one of the world's first commodity UDC smartphone prototypes to measure the real-world Point Spread Function (PSF) of the UDC system, and provide a model-based data synthesis pipeline to generate realistically degraded images. We specially design a new domain knowledge-enabled Dynamic Skip Connection Network (DISCNet) to restore the UDC images. We demonstrate the effectiveness of our method through extensive experiments on both synthetic and real UDC data. Our physics-based image formation model and proposed DISCNet can provide foundations for further exploration in UDC image restoration, and even for general diffraction artifact removal in a broader sense.
****
iVPF: Numerical Invertible Volume Preserving Flow for Efficient Lossless Compression
Shifeng Zhang, Chen Zhang, Ning Kang, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 620-629
It is nontrivial to store rapidly growing big data nowadays, which demands high-performance lossless compression techniques. Likelihood-based generative models have witnessed their success on lossless compression, where flow based models are desirable in allowing exact data likelihood optimisation with bijective mappin

gs. However, common continuous flows are in contradiction with the discreteness of coding schemes, which requires either 1) imposing strict constraints on flow models that degrades the performance or 2) coding numerous bijective mapping errors which reduces the efficiency. In this paper, we investigate volume preserving flows for lossless compression and show that a bijective mapping without error is possible. We propose Numerical Invertible Volume Preserving Flow (iVPF) which is derived from the general volume preserving flows. By introducing novel computation algorithms on flow models, an exact bijective mapping is achieved without any numerical error. We also propose a lossless compression algorithm based on iVPF. Experiments on various datasets show that the algorithm based on iVPF achieves state-of-the-art compression ratio over lightweight compression algorithms.
****
Pose Recognition With Cascade Transformers
Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1944-1953
In this paper, we present a regression-based pose recognition method using cascade Transformers. One way to categorize the existing approaches in this domain is to separate them into 1). heatmap-based and 2). regression-based. In general, heatmap-based methods achieve higher accuracy but are subject to various heuristic designs (not end-to-end mostly), whereas regression-based approaches attain relatively lower accuracy but they have less intermediate non-differentiable steps. Here we utilize the encoder-decoder structure in Transformers to perform regression-based person and keypoint detection that is general-purpose and requires less heuristic design compared with the existing approaches. We demonstrate the keypoint hypothesis (query) refinement process across different self-attention layers to reveal the recursive self-attention mechanism in Transformers. In the experiments, we report competitive results for pose recognition when compared with the competing regression-based methods.
****
Data-Uncertainty Guided Multi-Phase Learning for Semi-Supervised Object Detection
Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4568-4577
In this paper, we delve into semi-supervised object detection where unlabeled images are leveraged to break through the upper bound of fully-supervised object detection models. Previous semi-supervised methods based on pseudo labels are severely degenerated by noise and prone to overfit to noisy labels, thus are deficient in learning different unlabeled knowledge well. To address this issue, we propose a data-uncertainty guided multi-phase learning method for semi-supervised object detection. We comprehensively consider divergent types of unlabeled images according to their difficulty levels, utilize them in different phases and ensemble models from different phases together to generate ultimate results. Image uncertainty guided easy data selection and region uncertainty guided RoI Re-weighting are involved in multi-phase learning and enable the detector to concentrate on more certain knowledge. Through extensive experiments on PASCAL VOC and MS COCO, we demonstrate that our method behaves extraordinarily compared to baseline approaches and outperforms them by a large margin, more than 3% on VOC and 2% on COCO.
****
Prototype-Guided Saliency Feature Learning for Person Search
Hanjae Kim, Sunghun Joung, Ig-Jae Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4865-4874
Existing person search methods integrate person detection and re-identification (re-ID) module into a unified system. Though promising results have been achieved, the misalignment problem, which commonly occurs in person search, limits the discriminative feature representation for re-ID. To overcome this limitation, we

introduce a novel framework to learn the discriminative representation by utilizing prototype in OIM loss. Unlike conventional methods using prototype as a representation of person identity, we utilize it as guidance to allow the attention network to consistently highlight multiple instances across different poses. Moreover, we propose a new prototype update scheme with adaptive momentum to increase the discriminative ability across different instances. Extensive ablation experiments demonstrate that our method can significantly enhance the feature discriminative power, outperforming the state-of-the-art results on two person search benchmarks including CUHK-SYSU and PRW.
****
Contrastive Learning for Compact Single Image Dehazing
Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10551-10560
Single image dehazing is a challenging ill-posed problem due to the severe information degeneration. However, existing deep learning based dehazing methods only adopt clear images as positive samples to guide the training of dehazing network while negative information is unexploited. Moreover, most of them focus on strengthening the dehazing network with an increase of depth and width, leading to a significant requirement of computation and memory. In this paper, we propose a novel contrastive regularization (CR) built upon contrastive learning to exploit both the information of hazy images and clear images as negative and positive samples, respectively. CR ensures that the restored image is pulled to closer to the clear image and pushed to far away from the hazy image in the representation space. Furthermore, considering trade-off between performance and memory storage, we develop a compact dehazing network based on autoencoder-like (AE) framework. It involves an adaptive mixup operation and a dynamic feature enhancement module, which can benefit from preserving information flow adaptively and expanding the receptive field to improve the network's transformation capability, respectively. We term our dehazing network with autoencoder and contrastive regularization as AECR-Net. The extensive experiments on synthetic and real-world datasets demonstrate that our AECR-Net surpass the state-of-the-art approaches. The code is released in https://github.com/GlassyWu/AECR-Net.
****
I3Net: Implicit Instance-Invariant Network for Adapting One-Stage Object Detectors
Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12576-12585
Recent works on two-stage cross-domain detection have widely explored the local feature patterns to achieve more accurate adaptation results. These methods heavily rely on the region proposal mechanisms and ROI-based instance-level features to design fine-grained feature alignment modules with respect to the foreground objects. However, for one-stage detectors, it is hard or even impossible to obtain explicit instance-level features in the detection pipelines. Motivated by this, we propose an Implicit Instance-Invariant Network (I3Net), which is tailored for adapting one-stage detectors and implicitly learns instance-invariant features via exploiting the natural characteristics of deep features in different layers. Specifically, we facilitate the adaptation from three aspects: (1) Dynamic and Class-Balanced Reweighting (DCBR) strategy, which considers the coexistence of intra-domain and intra-class variations to assign larger weights to those sample-scarce categories and easy-to-adapt samples; (2) Category-aware Object Pattern Matching (COPM) module, which boosts the cross-domain foreground objects matching guided by the categorical information and suppresses the uninformative background features; (3) Regularized Joint Category Alignment (RJCA) module, which jointly enforces the category alignment at different domain-specific layers with a consistency regularization. Experiments reveal that I3Net exceeds the state-of-the-art performance on benchmark datasets.
****
Body Meshes as Points

Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 546-556
We consider the challenging multi-person 3D body mesh estimation task in this work. Existing methods are mostly two-stage based--one stage for person localization and the other stage for individual body mesh estimation, leading to redundant pipelines with high computation cost and degraded performance for complex scenes (e.g., occluded person instances). In this work, we present a single stage model, Body Meshes as Points (BMP), to simplify the pipeline and lift both efficiency and performance. In particular, BMP adopts a new method that represents multiple person instances as points in the spatial-depth space where each point is associated with one body mesh. Hinging on such representations, BMP can directly predict body meshes for multiple persons in a single stage by concurrently localizing person instance points and estimating the corresponding body meshes. To better reason about depth ordering of all the persons within the same scene, BMP designs a simple yet effective inter-instance ordinal depth loss to obtain depth-coherent body mesh estimation. BMP also introduces a novel keypoint-aware augmentation to enhance model robustness to occluded person instances. Comprehensive experiments on benchmarks Panoptic, MuPoTS-3D and 3DPW clearly demonstrate the state-of-the-art efficiency of BMP for multi-person body mesh estimation, together with outstanding accuracy. Code can be found at: https://github.com/jfzhang95/BMP.
****
Pixel-Aligned Volumetric Avatars
Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, Stephen Lombardi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11733-11742
Acquisition and rendering of photo-realistic human heads is a highly challenging research problem of particular importance for virtual telepresence. Currently, the highest quality is achieved by volumetric approaches trained in a person-specific manner on multi-view data. These models better represent fine structure, such as hair, compared to simpler mesh-based models. Volumetric models typically employ a global code to represent facial expressions, such that they can be driven by a small set of animation parameters. While such architectures achieve impressive rendering quality, they can not easily be extended to the multi-identity setting. In this paper, we devise a novel approach for predicting volumetric avatars of the human head given just a small number of inputs. We enable generalization across identities by a novel parameterization that combines neural radiance fields with local, pixel-aligned features extracted directly from the inputs, thus side-stepping the need for very deep or complex networks. Our approach is trained in an end-to-end manner solely based on a photometric re-rendering loss without requiring explicit 3D supervision. We demonstrate that our approach outperforms the existing state of the art in terms of quality and is able to generate faithful facial expressions in a multi-identity setting.
****
UC2: Universal Cross-Lingual Cross-Modal Vision-and-Language Pre-Training
Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, Jingjing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4155-4165
Vision-and-language pre-training has achieved impressive success in learning multimodal representations between vision and language. To generalize this success to non-English languages, we introduce UC^2, the first machine translation-augmented framework for cross-lingual cross-modal representation learning. To tackle the scarcity problem of multilingual captions for image datasets, we first augment existing English-only datasets with other languages via machine translation (MT). Then we extend the standard Masked Language Modeling and Image-Text Matching training objectives to multilingual setting, where alignment between different languages is captured through shared visual context (eg. using image as pivot). To facilitate the learning of a joint embedding space of images and all languages of interest, we further propose two novel pre-training tasks, namely Maksed R

egion-to-Token Modeling (MRTM) and Visual Translation Language Modeling (VTLM), leveraging MT-enhanced translated data. Evaluation on multilingual image-text re trieval and multilingual visual question answering benchmarks demonstrates that our proposed framework achieves new state of the art on diverse non-English benc hmarks while maintaining comparable performance to monolingual pre-trained model s on English tasks.
****
Generative PointNet: Deep Energy-Based Learning on Unordered Point Sets for 3D G eneration, Reconstruction and Classification
Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 , pp. 14976-14985
We propose a generative model of unordered point sets, such as point clouds, in the forms of an energy-based model, where the energy function is parameterized b y an input-permutation-invariant bottom-up neural network. The energy function l earns a coordinate encoding of each point and then aggregates all individual poi nt features into an energy for the whole point cloud. We show that our model can be derived from the discriminative PointNet. The model can be trained by MCMC-b ased maximum likelihood learning (as well as its variants), without the help of any assisting networks like those in GANs and VAEs. Unlike most point cloud gene rator that relies on hand-crafting distance metrics, our model does not rely on hand-crafting distance metric for the point cloud generation, because it synthes izes point clouds by matching observed examples in terms of statistical properti es defined by the energy function. Furthermore, we can learn a short-run MCMC to ward the energy-based model as a flow-like generator for point cloud reconstruct ion and interpolation. The learned point cloud representation can be useful for point cloud classification. Experiments demonstrate the advantages of the propos ed generative model of point clouds.
****
Blur, Noise, and Compression Robust Generative Adversarial Networks
Takuhiro Kaneko, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2021, pp. 13579-13589
Generative adversarial networks (GANs) have gained considerable attention owing to their ability to reproduce images. However, they can recreate training images faithfully despite image degradation in the form of blur, noise, and compressio n, generating similarly degraded images. To solve this problem, the recently pro posed noise robust GAN (NR-GAN) provides a partial solution by demonstrating the ability to learn a clean image generator directly from noisy images using a two -generator model comprising image and noise generators. However, its application is limited to noise, which is relatively easy to decompose owing to its additiv e and reversible characteristics, and its application to irreversible image degr adation, in the form of blur, compression, and combination of all, remains a cha llenge. To address these problems, we propose blur, noise, and compression robus t GAN (BNCR-GAN) that can learn a clean image generator directly from degraded i mages without knowledge of degradation parameters (e.g., blur kernel types, nois e amounts, or quality factor values). Inspired by NR-GAN, BNCR-GAN uses a multip le-generator model composed of image, blur-kernel, noise, and quality-factor gen erators. However, in contrast to NR-GAN, to address irreversible characteristics , we introduce masking architectures adjusting degradation strength values in a data-driven manner using bypasses before and after degradation. Furthermore, to suppress uncertainty caused by the combination of blur, noise, and compression, we introduce adaptive consistency losses imposing consistency between irreversib le degradation processes according to the degradation strengths. We demonstrate the effectiveness of BNCR-GAN through large-scale comparative studies on CIFAR-1 0 and a generality analysis on FFHQ. In addition, we demonstrate the applicabili ty of BNCR-GAN in image restoration.
****
Invisible Perturbations: Physical Adversarial Examples Exploiting the Rolling Sh utter Effect
Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, Earlence Fernandes;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2021, pp. 14666-14675
Physical adversarial examples for camera-based computer vision have so far been achieved through visible artifacts -- a sticker on a Stop sign, colorful borders around eyeglasses or a 3D printed object with a colorful texture. An implicit a ssumption here is that the perturbations must be visible so that a camera can se nse them. By contrast, we contribute a procedure to generate, for the first time , physical adversarial examples that are invisible to human eyes. Rather than mo difying the victim object with visible artifacts, we modify light that illuminat es the object. We demonstrate how an attacker can craft a modulated light signal that adversarially illuminates a scene and causes targeted misclassifications o n a state-of-the-art ImageNet deep learning model. Concretely, we exploit the ra diometric rolling shutter effect in commodity cameras to create precise striping patterns that appear on images. To human eyes, it appears like the object is il luminated, but the camera creates an image with stripes that will cause ML model s to output the attacker-desired classification. We conduct a range of simulatio n and physical experiments with LEDs, demonstrating targeted attack rates up to 84%.
****
Introvert: Human Trajectory Prediction via Conditional 3D Attention
Nasim Shafiee, Taskin Padir, Ehsan Elhamifar; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16815-16825
Predicting human trajectories is an important component of autonomous moving pla tforms, such as social robots and self-driving cars. Human trajectories are affe cted by both the physical features of the environment and social interactions wi th other humans. Despite recent surge of studies on human path prediction, most works focus on static scene information, therefore, cannot leverage the rich dyn amic visual information of the scene. In this work, we propose Introvert, a mode l which predicts human path based on his/her observed trajectory and the dynamic scene context, captured via a conditional 3D visual attention mechanism working on the input video. Introvert infers both environment constraints and social in teractions through observing the dynamic scene instead of communicating with oth er humans, hence, its computational cost is independent of how crowded the surro unding of a target human is. In addition, to focus on relevant interactions and constraints for each human, Introvert conditions its 3D attention model on the o bserved trajectory of the target human to extract and focus on relevant spatio-t emporal primitives. Our experiments on five publicly available datasets show tha t the Introvert improves the prediction errors of the state of the art.
****
Camouflaged Object Segmentation With Distraction Mining
Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, Deng-Ping Fan; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2021, pp. 8772-8781
Camouflaged object segmentation (COS) aims to identify objects that are "perfect ly" assimilate into their surroundings, which has a wide range of valuable appli cations. The key challenge of COS is that there exist high intrinsic similaritie s between the candidate objects and noise background. In this paper, we strive t o embrace challenges towards effective and efficient COS. To this end, we develo p a bio-inspired framework, termed Positioning and Focus Network (PFNet), which mimics the process of predation in nature. Specifically, our PFNet contains two key modules, i.e., the positioning module (PM) and the focus module (FM). The PM is designed to mimic the detection process in predation for positioning the pot ential target objects from a global perspective and the FM is then used to perfo rm the identification process in predation for progressively refining the coarse prediction via focusing on the ambiguous regions. Notably, in the FM, we develo p a novel distraction mining strategy for the distraction region discovery and r emoval, to benefit the performance of estimation. Extensive experiments demonstr ate that our PFNet runs in real-time (72 FPS) and significantly outperforms 18 c utting-edge models on three challenging benchmark datasets under four standard m etrics.

****
RfD-Net: Point Scene Understanding by Semantic Instance Reconstruction
Yinyu Nie, Ji Hou, Xiaoguang Han, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4608-4618
Semantic scene understanding from point clouds is particularly challenging as the points reflect only a sparse set of the underlying 3D geometry. Previous works often convert point cloud into regular grids (e.g. voxels or bird-eye view images), and resort to grid-based convolutions for scene understanding. In this work, we introduce RfD-Net that jointly detects and reconstructs dense object surfaces directly from raw point clouds. Instead of representing scenes with regular grids, our method leverages the sparsity of point cloud data and focuses on predicting shapes that are recognized with high objectness. With this design, we decouple the instance reconstruction into global object localization and local shape prediction. It not only eases the difficulty of learning 2-D manifold surfaces from sparse 3D space, the point clouds in each object proposal convey shape details that support implicit function learning to reconstruct any high-resolution surfaces. Our experiments indicate that instance detection and reconstruction present complementary effects, where the shape prediction head shows consistent effects on improving object detection with modern 3D proposal network backbones. The qualitative and quantitative evaluations further demonstrate that our approach consistently outperforms the state-of-the-arts and improves over 11 of mesh IoU in object reconstruction.
****
In the Light of Feature Distributions: Moment Matching for Neural Style Transfer
Nikolai Kalischek, Jan D. Wegner, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9382-9391
Style transfer aims to render the content of a given image in the graphical/artistic style of another image. The fundamental concept underlying Neural Style Transfer (NST) is to interpret style as a distribution in the feature space of a Convolutional Neural Network, such that a desired style can be achieved by matching its feature distribution. We show that most current implementations of that concept have important theoretical and practical limitations, as they only partially align the feature distributions. We propose a novel approach that matches the distributions more precisely, thus reproducing the desired style more faithfully, while still being computationally efficient. Specifically, we adapt the dual form of Central Moment Discrepancy, as recently proposed for domain adaptation, to minimize the difference between the target style and the feature distribution of the output image. The dual interpretation of this metric explicitly matches all higher-order centralized moments and is therefore a natural extension of existing NST methods that only take into account the first and second moments. Our experiments confirm that the strong theoretical properties also translate to visually better style transfer, and better disentangle style from semantic image content.
****
DOTS: Decoupling Operation and Topology in Differentiable Architecture Search
Yu-Chao Gu, Li-Juan Wang, Yun Liu, Yi Yang, Yu-Huan Wu, Shao-Ping Lu, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12311-12320
Differentiable Architecture Search (DARTS) has attracted extensive attention due to its efficiency in searching for cell structures. DARTS mainly focuses on the operation search and derives the cell topology from the operation weights. However, the operation weights can not indicate the importance of cell topology and result in poor topology rating correctness. To tackle this, we propose to Decouple the Operation and Topology Search (DOTS), which decouples the topology representation from operation weights and makes an explicit topology search. DOTS is achieved by introducing a topology search space that contains combinations of candidate edges. The proposed search space directly reflects the search objective and can be easily extended to support a flexible number of edges in the searched

cell. Existing gradient-based NAS methods can be incorporated into DOTS for furt
her improvement by the topology search. Considering that some operations (e.g.,
Skip-Connection) can affect the topology, we propose a group operation search sc
heme to preserve topology-related operations for a better topology search. The e
xperiments on CIFAR10/100 and ImageNet demonstrate that DOTS is an effective sol
ution for differentiable NAS. The code is released at https://github.com/guyucha
o/DOTS.
****
DriveGAN: Towards a Controllable High-Quality Neural Simulation
Seung Wook Kim, Jonah Philion, Antonio Torralba, Sanja Fidler; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p
p. 5820-5829
Realistic simulators are critical for training and verifying robotics systems. W
hile most of the contemporary simulators are hand-crafted, a scaleable way to bu
ild simulators is to use machine learning to learn how the environment behaves i
n response to an action, directly from data. In this work, we aim to learn to si
mulate a dynamic environment directly in pixel-space, by watching unannotated se
quences of frames and their associated action pairs. We introduce a novel high-q
uality neural simulator referred to as DriveGAN that achieves controllability by
 disentangling different components without supervision. In addition to steering
 controls, it also includes controls for sampling features of a scene, such as t
he weather as well as the location of non-player objects. Since DriveGAN is a fu
lly differentiable simulator, it further allows for re-simulation of a given vid
eo sequence, offering an agent to drive through a recorded scene again, possibly
 taking different actions. We train DriveGAN on multiple datasets, including 160
 hours of real-world driving data. We showcase that our approach greatly surpass
es the performance of previous data-driven simulators, and allows for new featur
es not explored before.
****
Style-Aware Normalized Loss for Improving Arbitrary Style Transfer
Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, Prem Natarajan; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2021, pp. 134-143
Neural Style Transfer (NST) has quickly evolved from single-style to infinite-st
yle models, also known as Arbitrary Style Transfer (AST). Although appealing res
ults have been widely reported in literature, our empirical studies on four well
-known AST approaches (GoogleMagenta, AdaIN, LinearTransfer, and SANet) show tha
t more than 50% of the time, AST stylized images are not acceptable to human use
rs, typically due to under- or over-stylization. We systematically study the cau
se of this imbalanced style transferability (IST) and propose a simple yet effec
tive solution to mitigate this issue. Our studies show that the IST issue is rel
ated to the conventional AST style loss, and reveal that the root cause is the e
qual weightage of training samples irrespective of the properties of their corre
sponding style images, which biases the model towards certain styles. Through in
vestigation of the theoretical bounds of the AST style loss, we propose a new lo
ss that largely overcomes IST. Theoretical analysis and experimental results val
idate the effectiveness of our loss, with over 80% relative improvement in style
 deception rate and 98% relatively higher preference in human evaluation.
****
Wide-Depth-Range 6D Object Pose Estimation in Space
Yinlin Hu, Sebastien Speierer, Wenzel Jakob, Pascal Fua, Mathieu Salzmann; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2021, pp. 15870-15879
6D pose estimation in space poses unique challenges that are not commonly encoun
tered in the terrestrial setting. One of the most striking differences is the la
ck of atmospheric scattering, allowing objects to be visible from a great distan
ce while complicating illumination conditions. Currently available benchmark dat
asets do not place a sufficient emphasis on this aspect and mostly depict the ta
rget in close proximity. Prior work tackling pose estimation under large scale v
ariations relies on a two-stage approach to first estimate scale, followed by po

se estimation on a resized image patch. We instead propose a single-stage hierar chical end-to-end trainable network that is more robust to scale variations. We demonstrate that it outperforms existing approaches not only on images synthesiz ed to resemble images taken in space but also on standard benchmarks.
****
Learning Salient Boundary Feature for Anchor-free Temporal Action Localization
Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Ji lin Li, Feiyue Huang, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2021, pp. 3320-3329
Temporal action localization is an important yet challenging task in video under standing. Typically, such a task aims at inferring both the action category and localization of the start and end frame for each action instance in a long, untr immed video. While most current models achieve good results by using pre-defined anchors and numerous actionness, such methods could be bothered with both large number of outputs and heavy tuning of locations and sizes corresponding to diff erent anchors. Instead, anchor-free methods is lighter, getting rid of redundant hyper-parameters, but gains few attention. In this paper, we propose the first purely anchor-free temporal localization method, which is both efficient and eff ective. Our model includes (i) an end-to-end trainable basic predictor, (ii) a s aliency-based refinement module to gather more valuable boundary features for ea ch proposal with a novel boundary pooling, and (iii) several consistency constra ints to make sure our model can find the accurate boundary given arbitrary propo sals. Extensive experiments show that our method beats all anchor-based and acti onness-guided methods with a remarkable margin on THUMOS14, achieving state-of-t he-art results, and comparable ones on ActivityNet v1.3. Our code will be made a vailable upon publication.
****
Monocular Depth Estimation via Listwise Ranking Using the Plackett-Luce Model
Julian Lienen, Eyke Hullermeier, Ralph Ewerth, Nils Nommensen; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 14595-14604
In many real-world applications, the relative depth of objects in an image is cr ucial for scene understanding. Recent approaches mainly tackle the problem of de pth prediction in monocular images by treating the problem as a regression task. Yet, being interested in an order relation in the first place, ranking methods suggest themselves as a natural alternative to regression, and indeed, ranking a pproaches leveraging pairwise comparisons as training information ("object A is closer to the camera than B") have shown promising performance on this problem. In this paper, we elaborate on the use of so-called listwise ranking as a genera lization of the pairwise approach. Our method is based on the Plackett-Luce (PL) model, a probability distribution on rankings, which we combine with a state-of -the-art neural network architecture and a simple sampling strategy to reduce tr aining complexity. Moreover, taking advantage of the representation of PL as a r andom utility model, the proposed predictor offers a natural way to recover (shi ft-invariant) metric depth information from ranking-only data provided at traini ng time. An empirical evaluation on several benchmark datasets in a "zero-shot" setting demonstrates the effectiveness of our approach compared to existing rank ing and regression methods.
****
Holistic 3D Scene Understanding From a Single Image With Implicit Representation
Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, Shuaicheng Li u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2021, pp. 8833-8842
We present a new pipeline for holistic 3D scene understanding from a single imag e, which could predict object shape, object pose and scene layout. As it is a hi ghly ill-posed problem, existing methods usually suffer from inaccurate estimati on of both shapes and layout especially for the cluttered scene due to the heavy occlusion between objects. We propose to utilize the latest deep implicit repre sentation to solve this challenge. We not only propose an image-based local stru ctured implicit network to improve the object shape estimation, but also refine

3D object pose and scene layout via an novel implicit scene graph neural network that exploits the implicit local object features. A novel physical violation loss is also proposed to avoid incorrect context between objects. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods in terms of object shape, scene layout estimation, and 3D object detection.
****
MultiBodySync: Multi-Body Segmentation and Motion Estimation via 3D Scan Synchronization
Jiahui Huang, He Wang, Tolga Birdal, Minhyuk Sung, Federica Arrigoni, Shi-Min Hu, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7108-7118
We present MultiBodySync, a novel, end-to-end trainable multi-body motion segmentation and rigid registration framework for multiple input 3D point clouds. The two non-trivial challenges posed by this multi-scan multibody setting that we investigate are: (i) guaranteeing correspondence and segmentation consistency across multiple input point clouds capturing different spatial arrangements of bodies or body parts; and (ii) obtaining robust motion-based rigid body segmentation applicable to novel object categories. We propose an approach to address these issues that incorporates spectral synchronization into an iterative deep declarative network, so as to simultaneously recover consistent correspondences as well as motion segmentation. At the same time, by explicitly disentangling the correspondence and motion segmentation estimation modules, we achieve strong generalizability across different object categories. Our extensive evaluations demonstrate that our method is effective on various datasets ranging from rigid parts in articulated objects to individually moving objects in a 3D scene, be it single-view or full point clouds.
****
Learning Optical Flow From a Few Matches
Shihao Jiang, Yao Lu, Hongdong Li, Richard Hartley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16592-16600
State-of-the-art neural network models for optical flow estimation require a dense correlation volume at high resolutions for representing per-pixel displacement. Although the dense correlation volume is informative for accurate estimation, its heavy computation and memory usage hinders the efficient training and deployment of the models. In this paper, we show that the dense correlation volume representation is redundant and accurate flow estimation can be achieved with only a fraction of elements in it. Based on this observation, we propose an alternative displacement representation, named Sparse Correlation Volume, which is constructed directly by computing the k closest matches in one feature map for each feature vector in the other feature map and stored in a sparse data structure. Experiments show that our method can reduce computational cost and memory use significantly and produce fine-structure motion, while maintaining high accuracy compared to previous approaches with dense correlation volumes.
****
Learnable Motion Coherence for Correspondence Pruning
Yuan Liu, Lingjie Liu, Cheng Lin, Zhen Dong, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3237-3246
Motion coherence is an important clue for distinguishing true correspondences from false ones. Modeling motion coherence on sparse putative correspondences is challenging due to their sparsity and uneven distributions. Existing works on motion coherence are sensitive to parameter settings and have difficulty in dealing with complex motion patterns. In this paper, we introduce a network called Laplacian Motion Coherence Network (LMCNet) to learn motion coherence property for correspondence pruning. We propose a novel formulation of fitting coherent motions with a smooth function on a graph of correspondences and show that this formulation allows a closed-form solution by graph Laplacian. This closed-form solution enables us to design a differentiable layer in a learning framework to capture global motion coherence from putative correspondences. The global motion cohere

nce is further combined with local coherence extracted by another local layer to robustly detect inlier correspondences. Experiments demonstrate that LMCNet has superior performances to the state of the art in relative camera pose estimation and correspondences pruning of dynamic scenes.
****
ManipulaTHOR: A Framework for Visual Object Manipulation
Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4497-4506
The domain of Embodied AI has recently witnessed substantial progress, particularly in navigating agents within their environments. These early successes have laid the building blocks for the community to tackle tasks that require agents to actively interact with objects in their environment. Object manipulation is an established research domain within the robotics community and poses several challenges including manipulator motion, grasping and long-horizon planning, particularly when dealing with oft-overlooked practical setups involving visually rich and complex scenes, manipulation using mobile agents (as opposed to tabletop manipulation), and generalization to unseen environments and objects. We propose a framework for object manipulation built upon the physics-enabled, visually rich AI2-THOR framework and present a new challenge to the Embodied AI community known as ArmPointNav. This task extends the popular point navigation task to object manipulation and offers new challenges including 3D obstacle avoidance, manipulating objects in the presence of occlusion, and multi-object manipulation that necessitates long term planning. Popular learning paradigms that are successful on PointNav challenges show promise, but leave a large room for improvement.
****
DeepI2P: Image-to-Point Cloud Registration via Deep Classification
Jiaxin Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15960-15969
This paper presents DeepI2P: a novel approach for cross-modality registration between an image and a point cloud. Given an image (e.g. from a rgb-camera) and a general point cloud (e.g. from a 3D Lidar scanner) captured at different locations in the same scene, our method estimates the relative rigid transformation between the coordinate frames of the camera and Lidar. Learning common feature descriptors to establish correspondences for the registration is inherently challenging due to the lack of appearance and geometric correlations across the two modalities. We circumvent the difficulty by converting the registration problem into a classification and inverse camera projection optimization problem. A classification neural network is designed to label whether the projection of each point in the point cloud is within or beyond the camera frustum. These labeled points are subsequently passed into a novel inverse camera projection solver to estimate the relative pose. Extensive experimental results on Oxford Robotcar and KITTI datasets demonstrate the feasibility of our approach. Our source code is available at https://github.com/lijx10/DeepI2P
****
Scene-Intuitive Agent for Remote Embodied Visual Grounding
Xiangru Lin, Guanbin Li, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7036-7045
Humans learn from life events to form intuitions towards the understanding of visual environments and languages. Envision that you are instructed by a high-level instruction, "Go to the bathroom in the master bedroom and replace the blue towel on the left wall", what would you possibly do to carry out the task? Intuitively, we comprehend the semantics of the instruction to form an overview of where a bathroom is and what a blue towel is in mind; then, we navigate to the target location by consistently matching the bathroom appearance in mind with the current scene. In this paper, we present an agent that mimics such human behaviors. Specifically, we focus on the Remote Embodied Visual Referring Expression in Real Indoor Environments task, called REVERIE, where an agent is asked to correctly localize a remote target object specified by a concise high-level natural language instruction, and propose a two-stage training pipeline. In the first stage,

we pre-train the agent with two cross-modal alignment sub-tasks, namely the Scene Grounding task and the Object Grounding task. The agent learns where to stop in the Scene Grounding task and what to attend to in the Object Grounding task respectively. Then, to generate action sequences, we propose a memory-augmented attentive action decoder to smoothly fuse the pre-trained vision and language representations with the agent's past memory experiences. Without bells and whistles, experimental results show that our method outperforms previous state-of-the-art(SOTA) significantly, demonstrating the effectiveness of our method.
****
Human-Like Controllable Image Captioning With Verb-Specific Semantic Roles
Long Chen, Zhihong Jiang, Jun Xiao, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16846-16856
Controllable Image Captioning (CIC) -- generating image descriptions following designated control signals -- has received unprecedented attention over the last few years. To emulate the human ability in controlling caption generation, current CIC studies focus exclusively on control signals concerning objective properties, such as contents of interest or descriptive patterns. However, we argue that almost all existing objective control signals have overlooked two indispensable characteristics of an ideal control signal: 1) Event-compatible: all visual contents referred to in a single sentence should be compatible with the described activity. 2) Sample-suitable: the control signals should be suitable for a specific image sample. To this end, we propose a new control signal for CIC: Verb-specific Semantic Roles (VSR). VSR consists of a verb and some semantic roles, which represents a targeted activity and the roles of entities involved in this activity. Given a designated VSR, we first train a grounded semantic role labeling (GSRL) model to identify and ground all entities for each role. Then, we propose a semantic structure planner (SSP) to learn human-like descriptive semantic structures. Lastly, we use a role-shift captioning model to generate the captions. Extensive experiments and ablations demonstrate that our framework can achieve better controllability than several strong baselines on two challenging CIC benchmarks. Besides, we can generate multi-level diverse captions easily. The code is available at: https://github.com/mad-red/VSR-guided-CIC.
****
Enhancing the Transferability of Adversarial Attacks Through Variance Tuning
Xiaosen Wang, Kun He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1924-1933
Deep neural networks are vulnerable to adversarial examples that mislead the models with imperceptible perturbations. Though adversarial attacks have achieved incredible success rates in the white-box setting, most existing adversaries often exhibit weak transferability in the black-box setting, especially under the scenario of attacking models with defense mechanisms. In this work, we propose a new method called variance tuning to enhance the class of iterative gradient based attack methods and improve their attack transferability. Specifically, at each iteration for the gradient calculation, instead of directly using the current gradient for the momentum accumulation, we further consider the gradient variance of the previous iteration to tune the current gradient so as to stabilize the update direction and escape from poor local optima. Empirical results on the standard ImageNet dataset demonstrate that our method could significantly improve the transferability of gradient-based adversarial attacks. Besides, our method could be used to attack ensemble models or be integrated with various input transformations. Incorporating variance tuning with input transformations on iterative gradient-based attacks in the multi-model setting, the integrated method could achieve an average success rate of 90.1% against nine advanced defense methods, improving the current best attack performance significantly by 85.1%. Code is available at https://github.com/JHL-HUST/VT.
****
HistoGAN: Controlling Colors of GAN-Generated and Real Images via Color Histograms
Mahmoud Afifi, Marcus A. Brubaker, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7941-79

While generative adversarial networks (GANs) can successfully produce high-quali ty images, they can be challenging to control. Simplifying GAN-based image gener ation is critical for their adoption in graphic design and artistic work. This g oal has led to significant interest in methods that can intuitively control the appearance of images generated by GANs. In this paper, we present HistoGAN, a co lor histogram-based method for controlling GAN-generated images' colors. We focu s on color histograms as they provide an intuitive way to describe image color w hile remaining decoupled from domain-specific semantics. Specifically, we introd uce an effective modification of the recent StyleGAN architecture [31] to contro l the colors of GAN-generated images specified by a target color histogram featu re. We then describe how to expand HistoGAN to recolor real images. For image re coloring, we jointly train an encoder network along with HistoGAN. The recolorin g model, ReHistoGAN, is an unsupervised approach trained to encourage the networ k to keep the original image's content while changing the colors based on the gi ven target histogram. We show that this histogram-based approach offers a better way to control GAN-generated and real images' colors while producing more compe lling results compared to existing alternative strategies.
****
BiCnet-TKS: Learning Efficient Spatial-Temporal Representation for Video Person Re-Identification
Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, Shiguang Shan; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2014-2023
In this paper, we present an efficient spatial-temporal representation for video person re-identification (reID). Firstly, we propose a Bilateral Complementary Network (BiCnet) for spatial complementarity modeling. Specifically, BiCnet cont ains two branches. Detail Branch processes frames at original resolution to pres erve the detailed visual clues, and Context Branch with a down-sampling strategy is employed to capture long-range contexts. On each branch, BiCnet appends mult iple parallel and diverse attention modules to discover divergent body parts for consecutive frames, so as to obtain an integral characteristic of target identi ty. Furthermore, a Temporal Kernel Selection (TKS) block is designed to capture short-term as well as long-term temporal relations by an adaptive mode. TKS can be inserted into BiCnet at any depth to construct BiCnet-TKS for spatial-tempora l modeling. Experimental results on multiple benchmarks show that BiCnet-TKS out performs state-of-the-arts with about 50% less computations. The source code is available at https://github.com/blue-blue272/BiCnet-TKS.
****
Probabilistic Model Distillation for Semantic Correspondence
Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 , pp. 7505-7514
Semantic correspondence is a fundamental problem in computer vision, which aims at establishing dense correspondences across images depicting different instance s under the same category. This task is challenging due to large intra-class var iations and a severe lack of ground truth. A popular solution is to learn corres pondences from synthetic data. However, because of the limited intra-class appea rance and background variations within synthetically generated training data, th e model's capability for handling "real" image pairs using such strategy is intr insically constrained. We address this problem with the use of a novel Probabili stic Model Distillation (PMD) approach which transfers knowledge learned by a pr obabilistic teacher model on synthetic data to a static student model with the u se of unlabeled real image pairs. A probabilistic supervision reweighting (PSR) module together with a confidence-aware loss (CAL) is used to mine the useful kn owledge and alleviate the impact of errors. Experimental results on a variety of benchmarks show that our PMD achieves state-of-the-art performance. To demonstr ate the generalizability of our approach, we extend PMD to incorporate stronger supervision for better accuracy -- the probabilistic teacher is trained with str onger key-point supervision. Again, we observe the superiority of our PMD. The e

xtensive experiments verify that PMD is able to infer more reliable supervision signals from the probabilistic teacher for representation learning and largely alleviate the influence of errors in pseudo labels.
****
OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets
Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7190-7199
We propose a novel framework for creating large-scale photorealistic datasets of indoor scenes, with ground truth geometry, material, lighting and semantics. Our goal is to make the dataset creation process widely accessible, allowing researchers to transform scans into datasets with highquality ground truth. We demonstrate our framework by creating a photorealistic synthetic version of the publicly available ScanNet dataset with consistent layout, semantic labels, high quality spatially-varying BRDF and complex lighting. We render photorealistic images, as well as complex spatially-varying lighting, including direct, indirect and visibility components. Such a dataset enables important applications in inverse rendering, scene understanding and robotics. We show that deep networks trained on the proposed dataset achieve competitive performance for shape, material and lighting estimation on real images, enabling photorealistic augmented reality applications, such as object insertion and material editing. We also show our semantic labels may be used for segmentation and multitask learning. Finally, we demonstrate that our framework may also be integrated with physics engines, to create virtual robotics environments with unique ground truth such as friction coefficients and correspondence to real scenes. The dataset and all the tools to create such datasets will be publicly released, enabling others in the community to easily build large-scale datasets of their own.
****
SSAN: Separable Self-Attention Network for Video Representation Learning
Xudong Guo, Xun Guo, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12618-12627
Self-attention has been successfully applied to video representation learning due to the effectiveness of modeling long range dependencies. Existing approaches build the dependencies merely by computing the pairwise correlations along spatial and temporal dimensions simultaneously. However, spatial correlations and temporal correlations represent different contextual information of scenes and temporal reasoning. Intuitively, learning spatial contextual information first will benefit temporal modeling. In this paper, we propose a separable self-attention (SSA) module, which models spatial and temporal correlations sequentially, so that spatial contexts can be efficiently used in temporal modeling. By adding SSA module into 2D CNN, we build a SSA network (SSAN) for video representation learning. On the task of video action recognition, our approach outperforms state-of-the-art methods on Something-Something and Kinetics-400 datasets. Our models often outperform counterparts with shallower network and less modality. We further verify the semantic learning ability of our method in visual-language task of video retrieval, which showcase the homogeneity of video representations and text embeddings. On MSR-VTT and Youcook2 datasets, video representations learnt by SSA significantly improve the state-of-the-art methods.
****
4D Panoptic LiDAR Segmentation
Mehmet Aygun, Aljosa Osep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5527-5537
Temporal semantic scene understanding is critical for self-driving cars or robots operating in dynamic environments. In this paper, we propose 4D panoptic LiDAR segmentation to assign a semantic class and a temporally-consistent instance ID to a sequence of 3D points. To this end, we present an approach and a novel evaluation metric. Our approach determines a semantic class for every point while m

odeling object instances as probability distributions in the 4D spatio-temporal domain. We process multiple point clouds in parallel and resolve point-to-instance associations, effectively alleviating the need for explicit temporal data association. Inspired by recent advances in benchmarking of multi-object tracking, we propose to adopt a new evaluation metric that separates the semantic and point-to-instance association aspects of the task. With this work, we aim at paving the road for future developments aiming at temporal LiDAR panoptic perception.
****
SceneGen: Learning To Generate Realistic Traffic Scenes
Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 892-901
We consider the problem of generating realistic traffic scenes automatically. Existing methods typically insert actors into the scene according to a set of hand-crafted heuristics and are limited in their ability to model the true complexity and diversity of real traffic scenes, thus inducing a content gap between synthesized traffic scenes versus real ones. As a result, existing simulators lack the fidelity necessary to train and test self-driving vehicles. To address this limitation, we present SceneGen, a neural autoregressive model of traffic scenes that eschews the need for rules and heuristics. In particular, given the ego-vehicle state and a high definition map of surrounding area, SceneGen inserts actors of various classes into the scene and synthesizes their sizes, orientations, and velocities. We demonstrate on two large-scale datasets SceneGen's ability to faithfully model distributions of real traffic scenes. Moreover, we show that SceneGen coupled with sensor simulation can be used to train perception models that generalize to the real world.
****
Natural Adversarial Examples
Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, Dawn Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15262-15271
We introduce two challenging datasets that reliably cause machine learning model performance to substantially degrade. The datasets are collected with a simple adversarial filtration technique to create datasets with limited spurious cues. Our datasets' real-world, unmodified examples transfer to various unseen models reliably, demonstrating that computer vision models have shared weaknesses. The first dataset is called ImageNet-A and is like the ImageNet test set, but it is far more challenging for existing models. We also curate an adversarial out-of-distribution detection dataset called ImageNet-O, which is the first out-of-distribution detection dataset created for ImageNet models. On ImageNet-A a DenseNet-121 obtains around 2% accuracy, an accuracy drop of approximately 90%, and its out-of-distribution detection performance on ImageNet-O is near random chance levels. We find that existing data augmentation techniques hardly boost performance, and using other public training datasets provides improvements that are limited. However, we find that improvements to computer vision architectures provide a promising path towards robust models.
****
CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models
Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, Jun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9593-9602
Learning disentanglement aims at finding a low dimensional representation which consists of multiple explanatory and generative factors of the observational data. The framework of variational autoencoder (VAE) is commonly used to disentangle independent factors from observations. However, in real scenarios, factors with semantics are not necessarily independent. Instead, there might be an underlying causal structure which renders these factors dependent. We thus propose a new VAE based framework named CausalVAE, which includes a Causal Layer to transform independent exogenous factors into causal endogenous ones that correspond to ca

usally related concepts in data. We further analyze the model identifiabitily, s
howing that the proposed model learned from observations recovers the true one u
p to a certain degree. Experiments are conducted on various datasets, including
synthetic and real word benchmark CelebA. Results show that the causal represent
ations learned by CausalVAE are semantically interpretable, and their causal rel
ationship as a Directed Acyclic Graph (DAG) is identified with good accuracy. Fu
rthermore, we demonstrate that the proposed CausalVAE model is able to generate
counterfactual data through "do-operation" to the causal factors.
****
VideoMoCo: Contrastive Video Representation Learning With Temporally Adversarial
 Examples
Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, Wei Liu; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1
1205-11214
MoCo is effective for unsupervised image representation learning. In this paper,
 we propose VideoMoCo for unsupervised video representation learning. Given a vi
deo sequence as an input sample, we improve the temporal feature representations
 of MoCo from two perspectives. First, we introduce a generator to drop out seve
ral frames from this sample temporally. The discriminator is then learned to enc
ode similar feature representations regardless of frame removals. By adaptively
dropping out different frames during training iterations of adversarial learning
, we augment this input sample to train a temporally robust encoder. Second, we
use temporal decay to model key attenuation in the memory queue when computing t
he contrastive loss. As the momentum encoder updates after keys enqueue, the rep
resentation ability of these keys degrades when we use the current input sample
for contrastive learning. This degradation is reflected via temporal decay to at
tend the input sample to recent keys in the queue. As a result, we adapt MoCo to
 learn video representations without empirically designing pretext tasks. By emp
owering the temporal robustness of the encoder and modeling the temporal decay o
f the keys, our VideoMoCo improves MoCo temporally based on contrastive learning
. Experiments on benchmark datasets including UCF101 and HMDB51 show that VideoM
oCo stands as a state-of-the-art video representation learning method.
****
Zero-Shot Instance Segmentation
Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, Li Cui; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 259
3-2602
Deep learning has significantly improved the precision of instance segmentation
with abundant labeled data. However, in many areas like medical and manufacturin
g, collecting sufficient data is extremely hard and labeling this data requires
high professional skills. We follow this motivation and propose a new task set n
amed zero-shot instance segmentation (ZSI). In the training phase of ZSI, the mo
del is trained with seen data, while in the testing phase, it is used to segment
 all seen and unseen instances. We first formulate the ZSI task and propose a me
thod to tackle the challenge, which consists of Zero-shot Detector, Semantic Mas
k Head, Background Aware RPN and Synchronized Background Strategy. We present a
new benchmark for zero-shot instance segmentation based on the MS-COCO dataset.
The extensive empirical results in this benchmark show that our method not only
surpasses the state-of-the-art results in zero-shot object detection task but al
so achieves promising performance on ZSI. Our approach will serve as a solid bas
eline and facilitate future research in zero-shot instance segmentation. Code av
ailable at ZSI.
****
Stereo Radiance Fields (SRF): Learning View Synthesis for Sparse Views of Novel
Scenes
Julian Chibane, Aayush Bansal, Verica Lazova, Gerard Pons-Moll; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
pp. 7911-7920
Recent neural view synthesis methods have achieved impressive quality and realis
m, surpassing classical pipelines which rely on multi-view reconstruction. State

-of-the-Art methods, such as NeRF, are designed to learn a single scene with a n
eural network and require dense multi-view inputs. Testing on a new scene requir
es re-training from scratch, which takes 2-3 days. In this work, we introduce St
ereo Radiance Fields (SRF), a neural view synthesis approach that is trained end
-to-end, generalizes to new scenes, and requires only sparse views at test time.
 The core idea is a neural architecture inspired by classical multi-view stereo
methods, which estimates surface points by finding similar image regions in ster
eo images. In SRF, we predict color and density for each 3D point given an encod
ing of its stereo correspondence in the input images. The encoding is implicitly
 learned by an ensemble of pair-wise similarities -- emulating classical stereo.
 Experiments show that SRF learns structure instead of overfitting on a scene. W
e train on multiple scenes of the DTU dataset and generalize to new ones without
 re-training, requiring only 10 sparse and spread-out views as input. We show th
at 10-15 minutes of fine-tuning further improve the results, achieving significa
ntly sharper, more detailed results than scene-specific models. The code, model,
 and videos are available at https://virtualhumans.mpi-inf.mpg.de/srf/.
****
Global Transport for Fluid Reconstruction With Learned Self-Supervision
Erik Franz, Barbara Solenthaler, Nils Thuerey; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1632-1642
We propose a novel method to reconstruct volumetric flows from sparse views via
a global transport formulation. Instead of obtaining the space-time function of
the observations, we reconstruct its motion based on a single initial state. In
addition we introduce a learned self-supervision that constrains observations fr
om unseen angles. These visual constraints are coupled via the transport constra
ints and a differentiable rendering step to arrive at a robust end-to-end recons
truction algorithm. This makes the reconstruction of highly realistic flow motio
ns possible, even from only a single input view. We show with a variety of synth
etic and real flows that the proposed global reconstruction of the transport pro
cess yields an improved reconstruction of the fluid motion.
****
SliceNet: Deep Dense Depth Estimation From a Single Indoor Panorama Using a Slic
e-Based Representation
Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, Enrico Gobbetti; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2021, pp. 11536-11545
We introduce a novel deep neural network to estimate a depth map from a single m
onocular indoor panorama. The network directly works on the equirectangular proj
ection, exploiting the properties of indoor 360 images. Starting from the fact t
hat gravity plays an important role in the design and construction of man-made i
ndoor scenes, we propose a compact representation of the scene into vertical sli
ces of the sphere, and we exploit long- and short-term relationships among slice
s to recover the equirectangular depth map. Our design makes it possible to main
tain high-resolution information in the extracted features even with a deep netw
ork. The experimental results demonstrate that our method outperforms current st
ate-of-the-art solutions in prediction accuracy, particularly for real-world dat
a.
****
Offboard 3D Object Detection From Point Cloud Sequences
Charles R. Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, Dragomir
Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2021, pp. 6134-6144
While current 3D object recognition research mostly focuses on the real-time, on
board scenario, there are many offboard use cases of perception that are largely
 under-explored, such as using machines to automatically generate high-quality 3
D labels. Existing 3D object detectors fail to satisfy the high-quality requirem
ent for offboard uses due to the limited input and speed constraints. In this pa
per, we propose a novel offboard 3D object detection pipeline using point cloud
sequence data. Observing that different frames capture complementary views of ob
jects, we design the offboard detector to make use of the temporal points throug

h both multi-frame object detection and novel object-centric refinement models. Evaluated on the Waymo Open Dataset, our pipeline named 3D Auto Labeling shows significant gains compared to the state-of-the-art onboard detectors and our offboard baselines. Its performance is even on par with human labels verified through a human label study. Further experiments demonstrate the application of auto labels for semi-supervised learning and provide extensive analysis to validate various design choices.
****
STaR: Self-Supervised Tracking and Reconstruction of Rigid Objects in Motion With Neural Rendering
Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, Steven Lovegrove; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13144-13152
We present STaR, a novel method that performs Self-supervised Tracking and Reconstruction of dynamic scenes with rigid motion from multi-view RGB videos without any manual annotation. Recent work has shown that neural networks are surprisingly effective at the task of compressing many views of a scene into a learned function which maps from a viewing ray to an observed radiance value via volume rendering. Unfortunately, these methods lose all their predictive power once any object in the scene has moved. In this work, we explicitly model rigid motion of objects in the context of neural representations of radiance fields. We show that without any additional human specified supervision, we can reconstruct a dynamic scene with a single rigid object in motion by simultaneously decomposing it into its two constituent parts and encoding each with its own neural representation. We achieve this by jointly optimizing the parameters of two neural radiance fields and a set of rigid poses which align the two fields at each frame. On both synthetic and real world datasets, we demonstrate that our method can render photorealistic novel views, where novelty is measured on both spatial and temporal axes. Our factored representation furthermore enables animation of unseen object motion.
****
Generalization on Unseen Domains via Inference-Time Label-Preserving Target Projections
Prashant Pandey, Mrigank Raman, Sumanth Varambally, Prathosh AP; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12924-12933
Generalization of machine learning models trained on a set of source domains on unseen target domains with different statistics, is a challenging problem. While many approaches have been proposed to solve this problem, they only utilize source data during training, but do not take advantage of the fact that a single target example is available at the time of inference. Motivated by this, we propose a method that effectively uses the target sample during inference beyond mere classification. Our method has three components - (i) A label preserving feature or metric transformation on source data such that the source samples are clustered in accordance with their class irrespective of their domain (ii) A generative model trained on the these features (iii) A label-preserving projection of the target point on the source-feature manifold during inference via solving an optimization problem on the input space of the generative model using the learned metric. Finally, the projected target is used in the classifier. Since the projected target feature comes from the source manifold and has the same label as the real target by design, the classifier is expected to perform better on it than the true target. We demonstrate that our method outperforms the state-of-the-art Domain Generalization methods on multiple datasets and tasks.
****
Monocular 3D Object Detection: An Extrinsic Parameter Free Approach
Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, Qinhong Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7556-7566
Monocular 3D object detection is an important task in autonomous driving. It can be easily intractable where there exists ego-car pose change w.r.t. ground plan

e. This is common due to the slight fluctuation of road smoothness and slope. Du
e to the lack of insight in industrial application, existing methods on open dat
asets neglect camera pose information, which inevitably results in the detector
being susceptible to camera extrinsic parameters. The perturbation of objects is
 very popular in most autonomous driving cases for industrial products. To this
end, we propose a novel method to capture camera pose to formulate the detector
free from extrinsic perturbation. Specifically, the proposed framework predicts
camera extrinsic parameters by detecting vanishing point and horizon change. A c
onverter is designed to rectify perturbative features in the latent space. By do
ing so, our 3D detector works independent of the extrinsic parameter variations,
 and produces accurate results in realistic cases, e.g., potholed and uneven roa
ds, where almost all existing monocular detectors fail to handle. Experiments de
monstrate our method yields best performance compared with the other state-of-th
e-arts by a large margin on both KITTI 3D and nuScenes datasets.
****
Communication Efficient SGD via Gradient Sampling With Bayes Prior
Liuyihan Song, Kang Zhao, Pan Pan, Yu Liu, Yingya Zhang, Yinghui Xu, Rong Jin; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2021, pp. 12065-12074
Gradient compression has been widely adopted in data-parallel distributed traini
ng of deep neural networks to reduce communication overhead. Some literatures ha
ve demonstrated that large gradients are more important than small ones because
they contain more information, such as Top-k compressor. Other mainstream method
s, like random-k compressor and gradient quantization, usually treat all gradien
ts equally. Different from all of them, we regard large and small gradients sele
ction as the exploitation and exploration of gradient information, respectively.
 And we find taking both of them into consideration is the key to boost the fina
l accuracy. So, we propose a novel gradient compressor: Gradient Sampling with B
ayes Prior in this paper. Specifically, we sample important/large gradients base
d on the global gradient distribution, which is periodically updated across mult
iple workers. Then we introduce Bayes Prior into distribution model to further e
xplore the gradients. We prove the convergence of our method for smooth non-conv
ex problems in the distributed system. Compared with methods that running after
high compression ratio at the expense of accuracy, we pursue no loss of accuracy
 and the actual acceleration benefit in practice. Experimental comparisons on a
variety of computer vision tasks (e.g. image classification and object detection
) and backbones (ResNet, MobileNetV2, InceptionV3 and AlexNet) show that our app
roach outperforms the state-of-the-art techniques in terms of both speed and acc
uracy, with the limitation of 100* compression ratio.
****
AdaBins: Depth Estimation Using Adaptive Bins
Shariq Farooq Bhat, Ibraheem Alhashim, Peter Wonka; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4009-401
8
We address the problem of estimating a high quality dense depth map from a singl
e RGB input image. We start out with a baseline encoder-decoder convolutional ne
ural network architecture and pose the question of how the global processing of
information can help improve overall depth estimation. To this end, we propose a
 transformer-based architecture block that divides the depth range into bins who
se center value is estimated adaptively per image. The final depth values are es
timated as linear combinations of the bin centers. We call our new building bloc
k AdaBins. Our results show a decisive improvement over the state-of-the-art on
several popular depth datasets across all metrics. We also validate the effectiv
eness of the proposed block with an ablation study and provide the code and corr
esponding pre-trained weights of the new state-of-the-art model.
****
VirFace: Enhancing Face Recognition via Unlabeled Shallow Data
Wenyu Li, Tianchu Guo, Pengyu Li, Binghui Chen, Biao Wang, Wangmeng Zuo, Lei Zha
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2021, pp. 14729-14738

Recently, exploiting the effect of the unlabeled data for face recognition attra
cts increasing attention. However, there are still few works considering the sit
uation that the unlabeled data is shallow which widely exists in real-world scen
arios. The existing semi-supervised face recognition methods which focus on gene
rating pseudo labels or minimizing softmax classification probabilities of the u
nlabeled data don't work very well on the unlabeled shallow data. Thus, it is st
ill a challenge on how to effectively utilize the unlabeled shallow face data fo
r improving the performance of face recognition. In this paper, we propose a nov
el face recognition method, named VirFace, to effectively apply the unlabeled sh
allow data for face recognition. VirFace consists of VirClass and VirInstance. S
pecifically, VirClass enlarges the inter-class distance by injecting the unlabel
ed data as new identities. Furthermore, VirInstance produces virtual instances s
ampled from the learned distribution of each identity to further enlarge the int
er-class distance. To the best of our knowledge, we are the first working on tac
kling the unlabeled shallow face data. Extensive experiments have been conducted
 on both the small- and large-scale datasets, e.g. LFW and IJB-C, etc, showing t
he superiority of the proposed method.
****

Pulsar: Efficient Sphere-Based Neural Rendering

Christoph Lassner, Michael Zollhofer; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1440-1449

We propose Pulsar, an efficient sphere-based differentiable rendering module tha
t is orders of magnitude faster than competing techniques, modular, and easy-to-
use due to its tight integration with PyTorch. Differentiable rendering is the f
oundation for modern neural rendering approaches, since it enables end-to-end tr
aining of 3D scene representations from image observations. However, gradient-ba
sed optimization of neural mesh, voxel, or function representations suffers from
 multiple challenges, i.e., topological in-consistencies, high memory footprints
, or slow rendering speeds. To alleviate these problems, Pulsar employs: 1) asph
ere-based scene representation, 2) a modular, efficient differentiable projectio
n operation, and 3) (optional) neural shading. Pulsar executes orders of magnitu
de faster than existing techniques and allows real-time rendering and optimizati
on of representations with millions of spheres. Using spheres for the scene repr
esentation, unprecedented speed is obtained while avoiding topology problems. Pu
lsar is fully differentiable and thus enables a plethora of applications, rangin
g from 3D reconstruction to neural rendering.
****

Contrastive Learning Based Hybrid Networks for Long-Tailed Image Classification

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, Lei Wang; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 943-9
52

Learning discriminative image representations plays a vital role in long-tailed
image classification because it can ease the classifier learning in imbalanced c
ases. Given the promising performance contrastive learning has shown recently in
 representation learning, in this work, we explore effective supervised contrast
ive learning strategies and tailor them to learn better image representations fr
om imbalanced data in order to boost the classification accuracy thereon. Specif
ically, we propose a novel hybrid network structure being composed of a supervis
ed contrastive loss to learn image representations and a cross-entropy loss to l
earn classifiers, where the learning is progressively transited from feature lea
rning to the classifier learning to embody the idea that better features make be
tter classifiers. We explore two variants of contrastive loss for feature learni
ng, which vary in the forms but share a common idea of pulling the samples from
the same class together in the normalized embedding space and pushing the sample
s from different classes apart. One of them is the recently proposed supervised
contrastive (SC) loss, which is designed on top of the state-of-the-art unsuperv
ised contrastive loss by incorporating positive samples from the same class. The
 other is a prototypical supervised contrastive (PSC) learning strategy which ad
dresses the intensive memory consumption in standard SC loss and thus shows more
 promise under limited memory budget. Extensive experiments on three long-tailed

classification datasets demonstrate the advantage of the proposed contrastive learning based hybrid networks in long-tailed classification.
****
Visualizing Adapted Knowledge in Domain Transfer
Yunzhong Hou, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13824-13833
A source model trained on source data and a target model learned through unsupervised domain adaptation (UDA) usually encode different knowledge. To understand the adaptation process, we portray their knowledge difference with image translation. Specifically, we feed a translated image and its original version to the two models respectively, formulating two branches. Through updating the translated image, we force similar outputs from the two branches. When such requirements are met, differences between the two images can compensate for and hence represent the knowledge difference between models. To enforce similar outputs from the two branches and depict the adapted knowledge, we propose a source-free image translation method that generates source-style images using only target images and the two models. We visualize the adapted knowledge on several datasets with different UDA methods and find that generated images successfully capture the style difference between the two domains. For application, we show that generated images enable further tuning of the target model without accessing source data. Code available at https://github.com/hou-yz/DA_visualization.
****
Delving into Data: Effectively Substitute Training for Black-box Attack
Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, Xiangyang Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4761-4770
Deep models have shown their vulnerability when processing adversarial samples. As for the black-box attack, without access to the architecture and weights of the attacked model, training a substitute model for adversarial attacks has attracted wide attention. Previous substitute training approaches focus on stealing the knowledge of the target model based on real training data or synthetic data, without exploring what kind of data can further improve the transferability between the substitute and target models. In this paper, we propose a novel perspective substitute training that focuses on designing the distribution of data used in the knowledge stealing process. More specifically, a diverse data generation module is proposed to synthesize large-scale data with wide distribution. And adversarial substitute training strategy is introduced to focus on the data distributed near the decision boundary. The combination of these two modules can further boost the consistency of the substitute model and target model, which greatly improves the effectiveness of adversarial attack. Extensive experiments demonstrate the efficacy of our method against state-of-the-art competitors under non-target and target attack settings. Detailed visualization and analysis are also provided to help understand the advantage of our method.
****
How To Exploit the Transferability of Learned Image Compression to Conventional Codecs
Jan P. Klopp, Keng-Chi Liu, Liang-Gee Chen, Shao-Yi Chien; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16165-16174
Lossy image compression is often limited by the simplicity of the chosen loss measure. Recent research suggests that generative adversarial networks have the ability to overcome this limitation and serve as a multi-modal loss, especially for textures. Together with learned image compression, these two techniques can be used to great effect when relaxing the commonly employed tight measures of distortion. However, convolutional neural network-based algorithms have a large computational footprint. Ideally, an existing conventional codec should stay in place, ensuring faster adoption and adherence to a balanced computational envelope. As a possible avenue to this goal, we propose and investigate how learned image coding can be used as a surrogate to optimise an image for encoding. A learned filter alters the image to optimise a different performance measure or a particul

ar task. Extending this idea with a generative adversarial network, we show how entire textures are replaced by ones that are less costly to encode but preserve a sense of detail. Our approach can remodel a conventional codec to adjust for the MS-SSIM distortion with over 20% rate improvement without any decoding overhead. On task-aware image compression, we perform favourably against a similar but codec-specific approach.
****
CorrNet3D: Unsupervised End-to-End Learning of Dense Correspondence for 3D Point Clouds
Yiming Zeng, Yue Qian, Zhiyu Zhu, Junhui Hou, Hui Yuan, Ying He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6052-6061
Motivated by the intuition that one can transform two aligned point clouds to each other more easily and meaningfully than a misaligned pair, we propose CorrNet3D -the first unsupervised and end-to-end deep learning-based framework - to drive the learning of dense correspondence between 3D shapes by means of deformation-like reconstruction to overcome the need for annotated data. Specifically, CorrNet3D consists of a deep feature embedding module and two novel modules called correspondence indicator and symmetric deformer. Feeding a pair of raw point clouds, our model first learns the pointwise features and passes them into the indicator to generate a learnable correspondence matrix used to permute the input pair. The symmetric deformer, with an additional regularized loss, transforms the two permuted point clouds to each other to drive the unsupervised learning of the correspondence. The extensive experiments on both synthetic and real-world datasets of rigid and non-rigid 3D shapes show our CorrNet3D outperforms state-of-the-art methods to a large extent, including those taking meshes as input. CorrNet3D is a flexible framework in that it can be easily adapted to supervised learning if annotated data are available. The source code and pre-trained model will be available at https://github.com/ZENGYIMINGEAMON/CorrNet3D.git.
****
Single-View Robot Pose and Joint Angle Estimation via Render & Compare
Yann Labbe, Justin Carpentier, Mathieu Aubry, Josef Sivic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1654-1663
We introduce RoboPose, a method to estimate the joint angles and the 6D camera-to-robot pose of a known articulated robot from a single RGB image. This is an important problem to grant mobile and itinerant autonomous systems the ability to interact with other robots using only visual information in non-instrumented environments, especially in the context of collaborative robotics. It is also challenging because robots have many degrees of freedom and an infinite space of possible configurations that often result in self-occlusions and depth ambiguities when imaged by a single camera. The contributions of this work are three-fold. First, we introduce a new render & compare approach for estimating the 6D pose and joint angles of an articulated robot that can be trained from synthetic data, generalizes to new unseen robot configurations at test time, and can be applied to a variety of robots. Second, we experimentally demonstrate the importance of the robot parametrization for the iterative pose updates and design a parametrization strategy that is independent of the robot structure. Finally, we show experimental results on existing benchmark datasets for four different robots and demonstrate that our method significantly outperforms the state of the art. Code and pre-trained models are available on the project webpage.
****
Harmonious Semantic Line Detection via Maximal Weight Clique Selection
Dongkwon Jin, Wonhui Park, Seong-Gyun Jeong, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16737-16745
A novel algorithm to detect an optimal set of semantic lines is proposed in this work. We develop two networks: selection network (S-Net) and harmonization network (H-Net). First, S-Net computes the probabilities and offsets of line candidates. Second, we filter out irrelevant lines through a selection-and-removal proc

ess. Third, we construct a complete graph, whose edge weights are computed by H-Net. Finally, we determine a maximal weight clique representing an optimal set of semantic lines. Moreover, to assess the overall harmony of detected lines, we propose a novel metric, called HIoU. Experimental results demonstrate that the proposed algorithm can detect harmonious semantic lines effectively and efficiently. Our codes are available at https://github.com/dongkwonjin/Semantic-Line-MWCS.
****
Learning the Non-Differentiable Optimization for Blind Super-Resolution
Zheng Hui, Jie Li, Xiumei Wang, Xinbo Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2093-2102
Previous convolutional neural network (CNN) based blind super-resolution (SR) methods usually adopt an iterative optimization way to approximate the ground-truth (GT) step-by-step. This solution always involves more computational costs to bring about time-consuming inference. At present, most blind SR algorithms are dedicated to obtaining high-fidelity results; their loss function generally employs L1 loss. To further improve the visual quality of SR results, perceptual metric, such as NIQE, is necessary to guide the network optimization. However, due to the non-differentiable property of NIQE, it cannot be as the loss function. Towards these issues, we propose an adaptive modulation network (AMNet) for multiple degradations SR, which is composed of the pivotal adaptive modulation layer (AMLayer). It is an efficient yet lightweight fusion layer between blur kernel and image features. Equipped with the blur kernel predictor, we naturally upgrade the AMNet to the blind SR model. Instead of considering iterative strategy, we make the blur kernel predictor trainable in the whole blind SR model, in which AMNet is well-trained. Also, we fit deep reinforcement learning into the blind SR model (AMNet-RL) to tackle the non-differentiable optimization problem. Specifically, the blur kernel predictor will be the actor to estimate the blur kernel from the input low-resolution (LR) image. The reward is designed by the pre-defined differentiable or non-differentiable metric. Extensive experiments show that our model can outperform state-of-the-art methods in both fidelity and perceptual metrics.
****
Progressive Temporal Feature Alignment Network for Video Inpainting
Xueyan Zou, Linjie Yang, Ding Liu, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16448-16457
Video inpainting aims to fill spatio-temporal "corrupted" regions with plausible content. To achieve this goal, it is necessary to find correspondences from neighbouring frames to faithfully hallucinate the unknown content. Current methods achieve this goal through attention, flow-based warping, or 3D temporal convolution. However, flow-based warping can create artifacts when optical flow is not accurate, while temporal convolution may suffer from spatial misalignment. We propose `Progressive Temporal Feature Alignment Network', which progressively enriches features extracted from the current frame with the feature warped from neighbouring frames using optical flow. Our approach corrects the spatial misalignment in the temporal feature propagation stage, greatly improving visual quality and temporal consistency of the inpainted videos. Using the proposed architecture, we achieve state-of-the-art performance on the DAVIS and FVI datasets compared to existing deep learning approaches. Code is available at https://github.com/MaureenZOU/TSAM.
****
Bottleneck Transformers for Visual Recognition
Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, Ashish Vaswani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16519-16529
We present BoTNet, a conceptually simple yet powerful backbone architecture that incorporates self-attention for multiple computer vision tasks including image classification, object detection and instance segmentation. By just replacing the spatial convolutions with global self-attention in the final three bottleneck blocks of a ResNet and no other changes, our approach improves upon the baseline

s significantly on instance segmentation and object detection while also reducin g the parameters, with minimal overhead in latency. Through the design of BoTNet , we also point out how ResNet bottleneck blocks with self-attention can be view ed as Transformer blocks. Without any bells and whistles, BoTNet achieves 44.4% Mask AP and 49.7% Box AP on the COCO Instance Segmentation benchmark using the M ask R-CNN framework; surpassing the previous best published single model and sin gle scale results of ResNeSt evaluated on the COCO validation set. Finally, we p resent a simple adaptation of the BoTNet design for image classification, result ing in models that achieve a strong performance of 84.7% top-1 accuracy on the I mageNet benchmark while being up to 1.64x faster in compute time than the popula r EfficientNet models on TPU-v3 hardware. We hope our simple and effective appro ach will serve as a strong baseline for future research in self-attention models for vision.
****
Calibrated RGB-D Salient Object Detection
Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, Li Cheng; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9471-9481
Complex backgrounds and similar appearances between objects and their surroundin gs are generally recognized as challenging scenarios in Salient Object Detection (SOD). This naturally leads to the incorporation of depth information in additi on to the conventional RGB image as input, known as RGB-D SOD or depth-aware SOD . Meanwhile, this emerging line of research has been considerably hindered by th e noise and ambiguity that prevail in raw depth images. To address the aforement ioned issues, we propose a Depth Calibration and Fusion (DCF) framework that con tains two novel components: 1) a learning strategy to calibrate the latent bias in the original depth maps towards boosting the SOD performance; 2) a simple yet effective cross reference module to fuse features from both RGB and depth modal ities. Extensive empirical experiments demonstrate that the proposed approach ac hieves superior performance against 27 state-of-the-art methods. Moreover, the p roposed depth calibration strategy as a preprocessing step, can be further appli ed to existing cutting-edge RGB-D SOD models and noticeable improvements are ach ieved.
****
S3: Neural Shape, Skeleton, and Skinning Fields for 3D Human Modeling
Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2021, pp. 13284-13293
Constructing and animating humans is an important component for building virtual worlds in a wide variety of applications such as virtual reality or robotics te sting in simulation. As there are exponentially many variations of humans with d ifferent shape, pose and clothing, it is critical to develop methods that can au tomatically reconstruct and animate humans at scale from real world data. Toward s this goal, we represent the pedestrian's shape, pose and skinning weights as n eural implicit functions that are directly learned from data. This representatio n enables us to handle a wide variety of different pedestrian shapes and poses w ithout explicitly fitting a human parametric body model, allowing us to handle a wider range of human geometries and topologies. We demonstrate the effectivenes s of our approach on various datasets and show that our reconstructions outperfo rm existing state-of-the-art methods. Furthermore, our re-animation experiments show that we can generate 3D human animations at scale from a single RGB image (and/or an optional LiDAR sweep) as input.
****
OSTeC: One-Shot Texture Completion
Baris Gecer, Jiankang Deng, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7628-7638
The last few years have witnessed the great success of non-linear generative mod els in synthesizing high-quality photorealistic face images. Many recent 3D faci al texture reconstruction and pose manipulation from a single image approaches s till rely on large and clean face datasets to train image-to-image Generative Ad

versarial Networks (GANs). Yet the collection of such a large scale high-resolution 3D texture dataset is still very costly and difficult to maintain age/ethnicity balance. Moreover, regression-based approaches suffer from generalization to the in-the-wild conditions and are unable to fine-tune to a target-image. In this work, we propose an unsupervised approach for one-shot 3D facial texture completion that does not require large-scale texture datasets, but rather harnesses the knowledge stored in 2D face generators. The proposed approach rotates an input image in 3D and fill-in the unseen regions by reconstructing the rotated image in a 2D face generator, based on the visible parts. Finally, we stitch the most visible textures at different angles in the UV image-plane. Further, we frontalize the target image by projecting the completed texture into the generator. The qualitative and quantitative experiments demonstrate that the completed UV textures and frontalized images are of high quality, resembles the original identity, can be used to train a texture GAN model for 3DMM fitting and improve pose-invariant face recognition.
****
Learning To Count Everything
Viresh Ranjan, Udbhav Sharma, Thu Nguyen, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3394-3403
Existing works on visual counting primarily focus on one specific category at a time, such as people, animals, and cells. In this paper, we are interested in counting everything, that is to count objects from any category given only a few annotated instances from that category. To this end, we pose counting as a few-shot regression task. To tackle this task, we present a novel method that takes a query image together with a few exemplar objects from the query image and predicts a density map for the presence of all objects of interest in the query image. We also present a novel adaptation strategy to adapt our network to any novel visual category at test time, using only a few exemplar objects from the novel category. We also introduce a dataset of 147 object categories containing over 6000 images that are suitable for the few-shot counting task. The images are annotated with two types of annotation, dots and bounding boxes, and they can be used for developing few-shot counting models. Experiments on this dataset shows that our method outperforms several state-of-the-art object detectors and few-shot counting approaches. Our code and dataset can be found at https://github.com/cvlab-stonybrook/LearningToCountEverything.
****
Robust Representation Learning With Feedback for Single Image Deraining
Chenghao Chen, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7742-7751
A deraining network can be interpreted as a conditional generator that aims at removing rain streaks from image. Most existing image deraining methods ignore model errors caused by uncertainty that reduces embedding quality. Unlike existing image deraining methods that embed low-quality features into the model directly, we replace low-quality features by latent high-quality features. The spirit of closed-loop feedback in the automatic control field is borrowed to obtain latent high-quality features. A new method for error detection and feature compensation is proposed to address model errors. Extensive experiments on benchmark datasets as well as specific real datasets demonstrate that the proposed method outperforms recent state-of-the-art methods. Code is available at: https://github.com/LI-Hao-SJTU/DerainRLNet
****
Fully Understanding Generic Objects: Modeling, Segmentation, and Reconstruction
Feng Liu, Luan Tran, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7423-7433
Inferring 3D structure of a generic object from a 2D image is a long-standing objective of computer vision. Conventional approaches either learn completely from CAD-generated synthetic data, which have difficulty in inference from real images, or generate 2.5D depth image via intrinsic decomposition, which is limited compared to the full 3D reconstruction. One fundamental challenge lies in how to

leverage numerous real 2D images without any 3D ground truth. To address this issue, we take an alternative approach with semi-supervised learning. That is, for a 2D image of a generic object, we decompose it into latent representations of category, shape and albedo, lighting and camera projection matrix, decode the representations to segmented 3D shape and albedo respectively, and fuse these components to render an image well approximating the input image. Using a category-adaptive 3D joint occupancy field (JOF), we show that the complete shape and albedo modeling enables us to leverage real 2D images in both modeling and model fitting. The effectiveness of our approach is demonstrated through superior 3D reconstruction from a single image, being either synthetic or real, and shape segmentation.
****
SSN: Soft Shadow Network for Image Compositing
Yichen Sheng, Jianming Zhang, Bedrich Benes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4380-4390
We introduce an interactive Soft Shadow Network (SSN) to generates controllable soft shadows for image compositing. SSN takes a 2D object mask as input and thus is agnostic to image types such as painting and vector art. An environment light map is used to control the shadow's characteristics, such as angle and softness. SSN employs an Ambient Occlusion Prediction module to predict an intermediate ambient occlusion map, which can be further refined by the user to provides geometric cues to modulate the shadow generation. To train our model, we design an efficient pipeline to produce diverse soft shadow training data using 3D object models. In addition, we propose an inverse shadow map representation to improve model training. We demonstrate that our model produces realistic soft shadows in real-time. Our user studies show that the generated shadows are often indistinguishable from shadows calculated by a physics-based renderer and users can easily use SSN through an interactive application to generate specific shadow effects in minutes.
****
MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection
Jia-Chang Feng, Fa-Ting Hong, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14009-14018
Weakly supervised video anomaly detection (WS-VAD) is to distinguish anomalies from normal events based on discriminative representations. Most existing works are limited in insufficient video representations. In this work, we develop a multiple instance self-training framework (MIST) to efficiently refine task-specific discriminative representations with only video-level annotations. In particular, MIST is composed of 1) a multiple instance pseudo label generator, which adapts a sparse continuous sampling strategy to produce more reliable clip-level pseudo labels, and 2) a self-guided attention boosted feature encoder that aims to automatically focus on anomalous regions in frames while extracting task-specific representations. Moreover, we adopt a self-training scheme to optimize both components and finally obtain a task-specific feature encoder. Extensive experiments on two public datasets demonstrate the efficacy of our method, and our method performs comparably or even better with existing supervised and weakly supervised methods, specifically obtaining a frame-level AUC 94.83% on ShanghaiTech.
****
VinVL: Revisiting Visual Representations in Vision-Language Models
Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5579-5588
This paper presents a detailed study of improving vision features and develops an improved object detection model for vision language (VL) tasks. Compared to the most widely used bottom-up and top-down model [2], the new model is bigger, pre-trained on much larger training corpora that combine multiple public annotated object detection datasets, and thus can generate representations of a richer collection of visual objects and concepts. While previous VL research focuses solely on improving the vision-language fusion model and leaves the object detection model improvement untouched, we present an empirical study to show that vision

features matter significantly in VL models. In our experiments we feed the visio n features generated by the new object detection model into a pre-trained transf ormer-based VL fusion model Oscar+, and fine-tune Oscar+ on a wide range of down stream VL tasks. Our results show that the new vision features significantly imp rove the performance across all VL tasks, creating new state-of-the-art results on seven public benchmarks. We will release the new object detection model to pu blic.
****
Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression
Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, Jingdong Wang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 14676-14686
In this paper, we are interested in the bottom-up paradigm of estimating human p oses from an image. We study the dense keypoint regression framework that is pre viously inferior to the keypoint detection and grouping framework. Our motivatio n is that regressing keypoint positions accurately needs to learn representation s that focus on the keypoint regions. We present a simple yet effective approach , named disentangled keypoint regression (DEKR). We adopt adaptive convolutions through pixel-wise spatial transformer to activate the pixels in the keypoint re gions and accordingly learn representations from them. We use a multi-branch str ucture for separate regression: each branch learns a representation with dedicat ed adaptive convolutions and regresses one keypoint. The resulting disentangled representations are able to attend to the keypoint regions, respectively, and th us the keypoint regression is spatially more accurate. We empirically show that the proposed direct regression method outperforms keypoint detection and groupin g methods and achieves superior bottom-up pose estimation results on two benchma rk datasets, COCO and CrowdPose. The code and models are available at https://gi thub.com/HRNet/DEKR.
****
CoMoGAN: Continuous Model-Guided Image-to-Image Translation
Fabio Pizzati, Pietro Cerri, Raoul de Charette; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14288-14298
CoMoGAN is a continuous GAN relying on the unsupervised reorganization of the ta rget data on a functional manifold. To that matter, we introduce a new Functiona l Instance Normalization layer and residual mechanism, which together disentangl e image content from position on target manifold. We rely on naive physics-inspi red models to guide the training while allowing private model/translations featu res. CoMoGAN can be used with any GAN backbone and allows new types of image tra nslation, such as cyclic image translation like timelapse generation, or detache d linear translation. On all datasets, it outperforms the literature. Our code i s available in this page: https://github.com/cv-rits/CoMoGAN.
****
Self-Supervised Video Hashing via Bidirectional Transformers
Shuyan Li, Xiu Li, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13549-13558
Most existing unsupervised video hashing methods are built on unidirectional mod els with less reliable training objectives, which underuse the correlations amon g frames and the similarity structure between videos. To enable efficient scalab le video retrieval, we propose a self-supervised video Hashing method based on B idirectional Transformers (BTH). Based on the encoder-decoder structure of trans formers, we design a visual cloze task to fully exploit the bidirectional correl ations between frames. To unveil the similarity structure between unlabeled vide o data, we further develop a similarity reconstruction task by establishing reli able and effective similarity connections in the video space. Furthermore, we de velop a cluster assignment task to exploit the structural statistics of the whol e dataset such that more discriminative binary codes can be learned. Extensive e xperiments implemented on three public benchmark datasets, FCVID, ActivityNet an d YFCC, demonstrate the superiority of our proposed approach.
****
From Synthetic to Real: Unsupervised Domain Adaptation for Animal Pose Estimatio

n
Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1482-1491
Animal pose estimation is an important field that has received increasing attention in the recent years. The main challenge for this task is the lack of labeled data. Existing works circumvent this problem with pseudo labels generated from data of other easily accessible domains such as synthetic data. However, these pseudo labels are noisy even with consistency check or confidence-based filtering due to the domain shift in the data. To solve this problem, we design a multi-scale domain adaptation module (MDAM) to reduce the domain gap between the synthetic and real data. We further introduce an online coarse-to-fine pseudo label updating strategy. Specifically, we propose a self-distillation module in an inner coarse-update loop and a mean-teacher in an outer fine-update loop to generate new pseudo labels that gradually replace the old ones. Consequently, our model is able to learn from the old pseudo labels at the early stage, and gradually switch to the new pseudo labels to prevent overfitting in the later stage. We evaluate our approach on the TigDog and VisDA 2019 datasets, where we outperform existing approaches by a large margin. We also demonstrate the generalization ability of our model by testing extensively on both unseen domains and unseen animal categories. Our code is available at the project website.
****
Safe Local Motion Planning With Self-Supervised Freespace Forecasting
Peiyun Hu, Aaron Huang, John Dolan, David Held, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12732-12741
Safe local motion planning for autonomous driving in dynamic environments requires forecasting how the scene evolves. Practical autonomy stacks adopt a semantic object-centric representation of a dynamic scene and build object detection, tracking, and prediction modules to solve forecasting. However, training these modules comes at an enormous human cost of manually annotated objects across frames. In this work, we explore future freespace as an alternative representation to support motion planning. Our key intuition is that it is important to avoid straying into occupied space regardless of what is occupying it. Importantly, computing ground-truth future freespace is annotation-free. First, we explore freespace forecasting as a self-supervised learning task. We then demonstrate how to use forecasted freespace to identify collision-prone plans from off-the-shelf motion planners. Finally, we propose future freespace as an additional source of annotation-free supervision. We demonstrate how to integrate such supervision into the learning-based planners. Experimental results on nuScenes and CARLA suggest both approaches lead to a significant reduction in collision rates.
****
Camera-Space Hand Mesh Recovery via Semantic Aggregation and Adaptive 2D-1D Registration
Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, Wen Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13274-13283
Recent years have witnessed significant progress in 3D hand mesh recovery. Nevertheless, because of the intrinsic 2D-to-3D ambiguity, recovering camera-space 3D information from a single RGB image remains challenging. To tackle this problem, we divide camera-space mesh recovery into two sub-tasks, i.e., root-relative mesh recovery and root recovery. First, joint landmarks and silhouette are extracted from a single input image to provide 2D cues for the 3D tasks. In the root-relative mesh recovery task, we exploit semantic relations among joints to generate a 3D mesh from the extracted 2D cues. Such generated 3D mesh coordinates are expressed relative to a root position, i.e., wrist of the hand. In the root recovery task, the root position is registered to the camera space by aligning the generated 3D mesh back to 2D cues, thereby completing camera-space 3D mesh recovery. Our pipeline is novel in that (1) it explicitly makes use of known semantic relations among joints and (2) it exploits 1D projections of the silhouette and mesh to achieve robust registration. Extensive experiments on popular datasets s

uch as FreiHAND, RHD, and Human3.6M demonstrate that our approach achieves state-of-the-art performance on both root-relative mesh recovery and root recovery. Our code is publicly available at https://github.com/SeanChenxy/HandMesh.
****

CondenseNet V2: Sparse Feature Reactivation for Deep Networks
Le Yang, Haojun Jiang, Ruojin Cai, Yulin Wang, Shiji Song, Gao Huang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3569-3578
Reusing features in deep networks through dense connectivity is an effective way to achieve high computational efficiency. The recent proposed CondenseNet has shown that this mechanism can be further improved if redundant features are removed. In this paper, we propose an alternative approach named sparse feature reactivation (SFR), aiming at actively increasing the utility of features for reusing. In the proposed network, named CondenseNetV2, each layer can simultaneously learn to 1) selectively reuse a set of most important features from preceding layers; and 2) actively update a set of preceding features to increase their utility for later layers. Our experiments show that the proposed models achieve promising performance on image classification (ImageNet and CIFAR) and object detection (MS COCO) in terms of both theoretical efficiency and practical speed.
****

Learning Graphs for Knowledge Transfer With Limited Labels
Pallabi Ghosh, Nirat Saini, Larry S. Davis, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11151-11161
Fixed input graphs are a mainstay in approaches that utilize Graph Convolution Networks (GCNs) for knowledge transfer. The standard paradigm is to utilize relationships in the input graph to transfer information using GCNs from training to testing nodes in the graph; for example, the semi-supervised, zero-shot, and few-shot learning setups. We propose a generalized framework for learning and improving the input graph as part of the standard GCN-based learning setup. Moreover, we use additional constraints between similar and dissimilar neighbors for each node in the graph by applying triplet loss on the intermediate layer output. We present results of semi-supervised learning on Citeseer, Cora, and Pubmed benchmarking datasets, and zero/few-shot action recognition on UCF101 and HMDB51 datasets, significantly outperforming current approaches. We also present qualitative results visualizing the graph connections that our approach learns to update.
****

DRANet: Disentangling Representation and Adaptation Networks for Unsupervised Cross-Domain Adaptation
Seunghun Lee, Sunghyun Cho, Sunghoon Im; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15252-15261
In this paper, we present DRANet, a network architecture that disentangles image representations and transfers the visual attributes in a latent space for unsupervised cross-domain adaptation. Unlike the existing domain adaptation methods that learn associated features sharing a domain, DRANet preserves the distinctiveness of each domain's characteristics. Our model encodes individual representations of content (scene structure) and style (artistic appearance) from both source and target images. Then, it adapts the domain by incorporating the transferred style factor into the content factor along with learnable weights specified for each domain. This learning framework allows bi-/multi-directional domain adaptation with a single encoder-decoder network and aligns their domain shift. Additionally, we propose a content-adaptive domain transfer module that helps retain scene structure while transferring style. Extensive experiments show our model successfully separates content-style factors and synthesizes visually pleasing domain-transferred images. The proposed method demonstrates state-of-the-art performance on standard digit classification tasks as well as semantic segmentation tasks.
****

Look Before You Leap: Learning Landmark Features for One-Stage Visual Grounding
Binbin Huang, Dongze Lian, Weixin Luo, Shenghua Gao; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16888-16897
An LBYL ('Look Before You Leap') Network is proposed for end-to-end trainable one-stage visual grounding. The idea behind LBYL-Net is intuitive and straightforward: we follow a language's description to localize the target object based on its relative spatial relation to 'Landmarks', which is characterized by some spatial positional words and some descriptive words about the object. The core of our LBYL-Net is a landmark feature convolution module that transmits the visual features with the guidance of linguistic description along with different directions. Consequently, such a module encodes the relative spatial positional relations between the current object and its context. Then we combine the contextual information from the landmark feature convolution module with the target's visual features for grounding. To make this landmark feature convolution light-weight, we introduce a dynamic programming algorithm (termed dynamic max pooling) with low complexity to extract the landmark feature. Thanks to the landmark feature convolution module, we mimic the human behavior of `Look Before You Leap` to design an LBYL-Net, which takes full consideration of contextual information. Extensive experiments show our method's effectiveness in four grounding datasets. Specifically, our LBYL-Net outperforms all state-of-the-art two-stage and one-stage methods on ReferitGame. On RefCOCO and RefCOCO+, Our LBYL-Net also achieves comparable results or even better results than existing one-stage methods. Code is available at https://github.com/svip-lab/LBYLNet.
****
Information Bottleneck Disentanglement for Identity Swapping
Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3404-3413
Improving the performance of face forgery detectors often requires more identity-swapped images of higher-quality. One core objective of identity swapping is to generate identity-discriminative faces that are distinct from the target while identical to the source. To this end, properly disentangling identity and identity-irrelevant information is critical and remains a challenging endeavor. In this work, we propose a novel information disentangling and swapping network, called InfoSwap, to extract the most expressive information for identity representation from a pre-trained face recognition model. The key insight of our method is to formulate the learning of disentangled representations as optimizing an information bottleneck trade-off, in terms of finding an optimal compression of the pre-trained latent features. Moreover, a novel identity contrastive loss is proposed for further disentanglement by requiring a proper distance between the generated identity and the target. While the most prior works have focused on using various loss functions to implicitly guide the learning of representations, we demonstrate that our model can provide explicit supervision for learning disentangled representations, achieving impressive performance in generating more identity-discriminative swapped faces.
****
DualGraph: A Graph-Based Method for Reasoning About Label Noise
HaiYang Zhang, XiMing Xing, Liang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9654-9663
Unreliable labels derived from large-scale dataset prevent neural networks from fully exploring the data. Existing methods of learning with noisy labels primarily take noise-cleaning-based and sample-selection-based methods. However, for numerous studies on account of the above two views, selected samples cannot take full advantage of all data points and cannot represent actual distribution of categories, in particular if label annotation is corrupted. In this paper, we start from a different perspective and propose a robust learning algorithm called DualGraph, which aims to capture structural relations among labels at two different levels with graph neural networks including instance-level and distribution-level relations. Specifically, the instance-level relation utilizes instance similarity characterize sample category, while the distribution-level relation describes instance similarity distribution from each sample to all other samples. Since

 the distribution-level relation is robust to label noise, our network propagate
s it as supervised signals to refine instance-level similarity. Combining two le
vel relations, we design an end-to-end training paradigm to counteract noisy lab
els while generating reliable predictions. We conduct extensive experiments on t
he noisy CIFAR-10 dataset, CIFAR-100 dataset, and the Clothing1M dataset. The re
sults demonstrate the advantageous performance of the proposed method in compari
son to state-of-the-art baselines.
****
Automatic Correction of Internal Units in Generative Neural Networks
Ali Tousi, Haedong Jeong, Jiyeon Han, Hwanil Choi, Jaesik Choi; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
pp. 7932-7940
Generative Adversarial Networks (GANs) have shown satisfactory performance in sy
nthetic image generation by devising complex network structure and adversarial t
raining scheme. Even though GANs are able to synthesize realistic images, there
exists a number of generated images with defective visual patterns which are kno
wn as artifacts. While most of the recent work tries to fix artifact generations
 by perturbing latent code, few investigate internal units of a generator to fix
 them. In this work, we devise a method that automatically identifies the intern
al units generating various types of artifact images. We further propose the seq
uential correction algorithm which adjusts the generation flow by modifying the
detected artifact units to improve the quality of generation while preserving th
e original outline. Our method outperforms the baseline method in terms of FID-s
core and shows satisfactory results with human evaluation.
****
Generating Manga From Illustrations via Mimicking Manga Creation Workflow
Lvmin Zhang, Xinrui Wang, Qingnan Fan, Yi Ji, Chunping Liu; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
5642-5651
We present a framework to generate manga from digital illustrations. In professi
onal mange studios, the manga create workflow consists of three key steps: (1) A
rtists use line drawings to delineate the structural outlines in manga storyboar
ds. (2) Artists apply several types of regular screentones to render the shading
, occlusion, and object materials. (3) Artists selectively paste irregular scree
n textures onto the canvas to achieve various background layouts or special effe
cts. Motivated by this workflow, we propose a data-driven framework to convert a
 digital illustration into three corresponding components: manga line drawing, r
egular screentone, and irregular screen texture. These components can be directl
y composed into manga images and can be further retouched for more plentiful man
ga creations. To this end, we create a large-scale dataset with these three comp
onents annotated by artists in a human-in-the-loop manner. We conduct both perce
ptual user study and qualitative evaluation of the generated manga, and observe
that our generated image layers for these three components are practically usabl
e in the daily works of manga artists. We provide 60 qualitative results and 15
additional comparisons in the supplementary material. We will make our presented
 manga dataset publicly available to assist related applications.
****
Multi-Decoding Deraining Network and Quasi-Sparsity Based Training
Yinglong Wang, Chao Ma, Bing Zeng; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2021, pp. 13375-13384
Existing deep deraining models are mainly learned via directly minimizing the st
atistical differences between rainy images and rain-free ground truths. They emp
hasize learning a mapping from rainy images to rain-free images with supervision
. Despite the demonstrated success, these methods do not perform well on restori
ng the fine-grained local details or removing blurry rainy traces. In this work,
 we aim to exploit the intrinsic priors of rainy images and develop intrinsic lo
ss functions to facilitate training deraining networks, which decompose a rainy
image into a rain-free background layer and a rainy layer containing intact rain
 streaks. To this end, we introduce the quasi-sparsity prior to train network so
 as to generate two sparse layers with intact textures of different objects. The

n we explore the low-value prior to compensate sparsity, forcing all rain streaks to enter into one layer while non-rain contents into another layer to restore image details. We introduce a multi-decoding structure to specially supervise the generation of multi-type deraining features. This helps to learn the most contributory features to deraining in respective spaces. Moreover, our model stabilizes the feature values from multi-spaces via information sharing to alleviate potential artifacts, which also accelerates the running speed. Extensive experiments show that the proposed deraining method outperforms the state-of-the-art approaches in terms of effectiveness and efficiency.
****
Open-Vocabulary Object Detection Using Captions
Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14393-14402
Despite the remarkable accuracy of deep neural networks in object detection, they are costly to train and scale due to supervision requirements. Particularly, learning more object categories typically requires proportionally more bounding box annotations. Weakly supervised and zero-shot learning techniques have been explored to scale object detectors to more categories with less supervision, but they have not been as successful and widely adopted as supervised models. In this paper, we put forth a novel formulation of the object detection problem, namely open-vocabulary object detection, which is more general, more practical, and more effective than weakly supervised and zero-shot approaches. We propose a new method to train object detectors using bounding box annotations for a limited set of object categories, as well as image-caption pairs that cover a larger variety of objects at a significantly lower cost. We show that the proposed method can detect and localize objects for which no bounding box annotation is provided during training, at a significantly higher accuracy than zero-shot approaches. Meanwhile, objects with bounding box annotation can be detected almost as accurately as supervised methods, which is significantly better than weakly supervised baselines. Accordingly, we establish a new state of the art for scalable object detection.
****
Unveiling the Potential of Structure Preserving for Weakly Supervised Object Localization
Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11642-11651
Weakly supervised object localization (WSOL) remains an open problem due to the deficiency of finding object extent information using a classification network. While prior works struggle to localize objects by various spatial regularization strategies, we argue that how to extract object structural information from the trained classification network is neglected. In this paper, we propose a two-stage approach, termed structure-preserving activation (SPA), towards fully leveraging the structure information incorporated in convolutional features for WSOL. In the first stage, a restricted activation module (RAM) is designed to alleviate the structure-missing issue caused by the classification network, based on the observation that the unbounded classification map and global average pooling layer drive the network to focus only on object parts. In the second stage, we propose a post-process approach, termed the self-correlation map generating (SCG) module to obtain structure-preserving localization maps on the basis of the activation maps acquired from the first stage. Specifically, we utilize the high-order self-correlation (HSC) to extract the inherent structural information retained in the learned model and then aggregate HSC of multiple points for precise object localization. Extensive experiments on two publicly available benchmarks including CUB-200-2011 and ILSVRC show that the proposed SPA achieves substantial and consistent performance gains compared with baseline approaches.
****
From Points to Multi-Object 3D Reconstruction
Francis Engelmann, Konstantinos Rematas, Bastian Leibe, Vittorio Ferrari; Procee

dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4588-4597
We propose a method to detect and reconstruct multiple 3D objects from a single RGB image. The key idea is to optimize for detection, alignment and shape jointly over all objects in the RGB image, while focusing on realistic and physically plausible reconstructions. To this end, we propose a key-point detector that localizes objects as center points and directly predicts all object properties, including 9-DoF bounding boxes and 3D shapes, all in a single forward pass. The method formulates 3D shape reconstruction as a shape selection problem, i.e. it selects among exemplar shapes from a given database. This makes it agnostic to shape representations, which enables a lightweight reconstruction of realistic and visually-pleasing shapes based on CAD-models, while the training objective is formulated around point clouds and voxel representations. A collision-loss promotes non-intersecting objects, further increasing the reconstruction realism. Given the RGB image, the presented approach performs lightweight reconstruction in a single-stage, it is real-time capable, fully differentiable and end-to-end trainable. Our experiments compare multiple approaches for 9-DoF bounding box estimation, evaluate the novel shape-selection mechanism and compare to recent methods in terms of 3D bounding box estimation and 3D shape reconstruction quality.
****
Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning
Bin Li, Yin Li, Kevin W. Eliceiri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14318-14328
We address the challenging problem of whole slide image (WSI) classification. WSIs have very high resolutions and usually lack localized annotations. WSI classification can be cast as a multiple instance learning (MIL) problem when only slide-level labels are available. We propose a MIL-based method for WSI classification and tumor detection that does not require localized annotations. Our method has three major components. First, we introduce a novel MIL aggregator that models the relations of the instances in a dual-stream architecture with trainable distance measurement. Second, since WSIs can produce large or unbalanced bags that hinder the training of MIL models, we propose to use self-supervised contrastive learning to extract good representations for MIL and alleviate the issue of prohibitive memory cost for large bags. Third, we adopt a pyramidal fusion mechanism for multiscale WSI features, and further improve the accuracy of classification and localization. Our model is evaluated on two representative WSI datasets. The classification accuracy of our model compares favorably to fully-supervised methods, with less than 2% accuracy gap across datasets. Our results also outperform all previous MIL-based methods. Additional benchmark results on standard MIL datasets further demonstrate the superior performance of our MIL aggregator on general MIL problems.
****
Regressive Domain Adaptation for Unsupervised Keypoint Detection
Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, Mingsheng Long; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6780-6789
Domain adaptation (DA) aims at transferring knowledge from a labeled source domain to an unlabeled target domain. Though many DA theories and algorithms have been proposed, most of them are tailored into classification settings and may fail in regression tasks, especially in the practical keypoint detection task. To tackle this difficult but significant task, we present a method of regressive domain adaptation (RegDA) for unsupervised keypoint detection. Inspired by the latest theoretical work, we first utilize an adversarial regressor to maximize the disparity on the target domain and train a feature generator to minimize this disparity. However, due to the high dimension of the output space, this regressor fails to detect samples that deviate from the support of the source. To overcome this problem, we propose two important ideas. First, based on our observation that the probability density of the output space is sparse, we introduce a spatial probability distribution to describe this sparsity and then use it to guide the

learning of the adversarial regressor. Second, to alleviate the optimization difficulty in the high-dimensional space, we innovatively convert the minimax game in the adversarial training to the minimization of two opposite goals. Extensive experiments show that our method brings large improvement by 8% to 11% in terms of PCK on different datasets.
****
Mask Guided Matting via Progressive Refinement Network
Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, Alan Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1154-1163
We propose Mask Guided (MG) Matting, a robust matting framework that takes a general coarse mask as guidance. MG Matting leverages a network (PRN) design which encourages the matting model to provide self-guidance to progressively refine the uncertain regions through the decoding process. A series of guidance mask perturbation operations are also introduced in the training to further enhance its robustness to external guidance. We show that PRN can generalize to unseen types of guidance masks such as trimap and low-quality alpha matte, making it suitable for various application pipelines. In addition, we revisit the foreground color prediction problem for matting and propose a surprisingly simple improvement to address the dataset issue. Evaluation on real and synthetic benchmarks shows that MG Matting achieves state-of-the-art performance using various types of guidance inputs. Code and models are available at https://github.com/yucornetto/MGMatting.
****
Monocular Reconstruction of Neural Face Reflectance Fields
Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4791-4800
The reflectance field of a face describes the reflectance properties responsible for complex lighting effects including diffuse, specular, inter-reflection and self shadowing. Most existing methods for estimating the face reflectance from a monocular image assume faces to be diffuse with very few approaches adding a specular component. This still leaves out important perceptual aspects of reflectance such as higher-order global illumination effects and self-shadowing. We present a new neural representation for face reflectance where we can estimate all components of the reflectance responsible for the final appearance from a monocular image. Instead of modeling each component of the reflectance separately using parametric models, our neural representation allows us to generate a basis set of faces in a geometric deformation-invariant space, parameterized by the input light direction, viewpoint and face geometry. We learn to reconstruct this reflectance field of a face just from a monocular image, which can be used to render the face from any viewpoint in any light condition. Our method is trained on a light-stage dataset, which captures 300 people illuminated with 150 light conditions from 8 viewpoints. We show that our method outperforms existing monocular reflectance reconstruction methods due to better capturing of physical effects, such as sub-surface scattering, specularities, self-shadows and other higher-order effects.
****
SelfSAGCN: Self-Supervised Semantic Alignment for Graph Convolution Network
Xu Yang, Cheng Deng, Zhiyuan Dang, Kun Wei, Junchi Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16775-16784
Graph convolution networks (GCNs) are a powerful deep learning approach and have been successfully applied to representation learning on graphs in a variety of real-world applications. Despite their success, two fundamental weaknesses of GCNs limit their ability to represent graph-structured data: poor performance when labeled data are severely scarce and indistinguishable features when more layers are stacked. In this paper, we propose a simple yet effective Self-Supervised Semantic Alignment Graph Convolution Network (SelfSAGCN), which consists of two

crux techniques: Identity Aggregation and Semantic Alignment, to overcome these weaknesses. The behind basic idea is the node features in the same class but lea rned from semantic and graph structural aspects respectively, are expected to be mapped nearby. Specifically, the Identity Aggregation is applied to extract sem antic features from labeled nodes, the Semantic Alignment is utilized to align n ode features obtained from different aspects using the class central similarity. In this way, the over-smoothing phenomenon is alleviated, while the similaritie s between the unlabeled features and labeled ones from the same class are enhanc ed. Experimental results on five popular datasets show that the proposed SelfSAG CN outperforms state-of-the-art methods on various classification tasks.
****
ECKPN: Explicit Class Knowledge Propagation Network for Transductive Few-Shot Le arning
Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, Zhe Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6596-6605
Recently, the transductive graph-based methods have achieved great success in th e few-shot classification task. However, most existing methods ignore exploring the class-level knowledge that can be easily learned by humans from just a handf ul of samples. In this paper, we propose an Explicit Class Knowledge Propagation Network (ECKPN), which is composed of the comparison, squeeze and calibration m odules, to address this problem. Specifically, we first employ the comparison mo dule to explore the pairwise sample relations to learn rich sample representatio ns in the instance-level graph. Then, we squeeze the instance-level graph to gen erate the class-level graph, which can help obtain the class-level visual knowle dge and facilitate modeling the relations of different classes. Next, the calibr ation module is adopted to characterize the relations of the classes explicitly to obtain the more discriminative class-level knowledge representations. Finally , we combine the class-level knowledge with the instance-level sample representa tions to guide the inference of the query samples. We conduct extensive experime nts on four few-shot classification benchmarks, and the experimental results sho w that the proposed ECKPN significantly outperforms the state-of-the-art methods .
****
Coarse-Fine Networks for Temporal Activity Detection in Videos
Kumara Kahatapitiya, Michael S. Ryoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8385-8394
In this paper, we introduce 'Coarse-Fine Networks', a two-stream architecture wh ich benefits from different abstractions of temporal resolution to learn better video representations for long-term motion. Traditional Video models process inp uts at one (or few) fixed temporal resolution without any dynamic frame selectio n. However, we argue that, processing multiple temporal resolutions of the input and doing so dynamically by learning to estimate the importance of each frame c an largely improve video representations, specially in the domain of temporal ac tivity localization. To this end, we propose (1) 'Grid Pool', a learned temporal downsampling layer to extract coarse features, and, (2) 'Multi-stage Fusion', a spatio-temporal attention mechanism to fuse a fine-grained context with the coa rse features. We show that our method outperforms the state-of-the-arts for acti on detection in public datasets including Charades with a significantly reduced compute and memory footprint. The code is available at https://github.com/kkahat apitiya/Coarse-Fine-Networks.
****
Can Audio-Visual Integration Strengthen Robustness Under Multimodal Attacks?
Yapeng Tian, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2021, pp. 5601-5611
In this paper, we propose to make a systematic study on machines' multisensory p erception under attacks. We use the audio-visual event recognition task against multimodal adversarial attacks as a proxy to investigate the robustness of audio -visual learning. We attack audio, visual, and both modalities to explore whethe r audio-visual integration still strengthens perception and how different fusion

mechanisms affect the robustness of audio-visual models. For interpreting the multimodal interactions under attacks, we learn a weakly-supervised sound source visual localization model to localize sounding regions in videos. To mitigate multimodal attacks, we propose an audio-visual defense approach based on an audio-visual dissimilarity constraint and external feature memory banks. Extensive experiments demonstrate that audio-visual models are susceptible to multimodal adversarial attacks; audio-visual integration could decrease the model robustness rather than strengthen under multimodal attacks; even a weakly-supervised sound source visual localization model can be successfully fooled; our defense method can improve the invulnerability of audio-visual networks without significantly sacrificing clean model performance.
****
Deep Gradient Projection Networks for Pan-sharpening
Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, Chunxia Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1366-1375
Pan-sharpening is an important technique for remote sensing imaging systems to obtain high resolution multispectral images. Recently, deep learning has become the most popular tool for pan-sharpening. This paper develops a model-based deep pan-sharpening approach. Specifically, two optimization problems regularized by the deep prior are formulated, and they are separately responsible for the generative models for panchromatic images and low resolution multispectral images. Then, the two problems are solved by a gradient projection algorithm, and the iterative steps are generalized into two network blocks. By alternatively stacking the two blocks, a novel network, called gradient projection based pan-sharpening neural network, is constructed. The experimental results on different kinds of satellite datasets demonstrate that the new network outperforms state-of-the-art methods both visually and quantitatively. The codes are available at https://github.com/xsxjtu/GPPNN.
****
ReNAS: Relativistic Evaluation of Neural Architecture Search
Yixing Xu, Yunhe Wang, Kai Han, Yehui Tang, Shangling Jui, Chunjing Xu, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4411-4420
An effective and efficient architecture performance evaluation scheme is essential for the success of Neural Architecture Search (NAS). To save computational cost, most of existing NAS algorithms often train and evaluate intermediate neural architectures on a small proxy dataset with limited training epochs. But it is difficult to expect an accurate performance estimation of an architecture in such a coarse evaluation way. This paper advocates a new neural architecture evaluation scheme, which aims to determine which architecture would perform better instead of accurately predict the absolute architecture performance. Therefore, we propose a relativistic architecture performance predictor in NAS (ReNAS). We encode neural architectures into feature tensors, and further refining the representations with the predictor. The proposed relativistic performance predictor can be deployed in discrete searching methods to search for the desired architectures without additional evaluation. Experimental results on NAS-Bench-101 dataset suggests that, sampling 424 (0.1% of the entire search space) neural architectures and their corresponding validation performance is already enough for learning an accurate architecture performance predictor. The accuracies of our searched neural architectures on NAS-Bench-101 and NAS-Bench-201 datasets are higher than that of the state-of-the-art methods and show the priority of the proposed method.
****
When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks
Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11855-11864
Human pose estimation is a fundamental yet challenging task in computer vision,

which aims at localizing human anatomical keypoints. However, unlike human visio
n that is robust to various data corruptions such as blur and pixelation, curren
t pose estimators are easily confused by these corruptions. This work comprehens
ively studies and addresses this problem by building rigorous robust benchmarks,
 termed COCO-C, MPII-C, and OCHuman-C, to evaluate the weaknesses of current adv
anced pose estimators, and a new algorithm termed AdvMix is proposed to improve
their robustness in different corruptions. Our work has several unique benefits.
 (1) AdvMix is model-agnostic and capable in a wide-spectrum of pose estimation
models. (2) AdvMix consists of adversarial augmentation and knowledge distillati
on. Adversarial augmentation contains two neural network modules that are traine
d jointly and competitively in an adversarial manner, where a generator network
mixes different corrupted images to confuse a pose estimator, improving the robu
stness of the pose estimator by learning from harder samples. To compensate for
the noise patterns by adversarial augmentation, knowledge distillation is applie
d to transfer clean pose structure knowledge to the target pose estimator. (3) E
xtensive experiments show that AdvMix significantly increases the robustness of
pose estimations across a wide range of corruptions, while maintaining accuracy
on clean data in various challenging benchmark datasets.
****
ReMix: Towards Image-to-Image Translation With Limited Data
Jie Cao, Luanxuan Hou, Ming-Hsuan Yang, Ran He, Zhenan Sun; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
15018-15027
Image-to-image (I2I) translation methods based on generative adversarial network
s (GANs) typically suffer from overfitting when limited training data is availab
le. In this work, we propose a data augmentation method (ReMix) to tackle this i
ssue. We interpolate training samples at the feature level and propose a novel c
ontent loss based on the perceptual relations among samples. The generator learn
s to translate the in-between samples rather than memorizing the training set, a
nd thereby forces the discriminator to generalize. The proposed approach effecti
vely reduces the ambiguity of generation and renders content-preserving results.
 The ReMix method can be easily incorporated into existing GAN models with minor
 modifications. Experimental results on numerous tasks demonstrate that GAN mode
ls equipped with the ReMix method achieve significant improvements.
****
Adaptive Rank Estimate in Robust Principal Component Analysis
Zhengqin Xu, Rui He, Shoulie Xie, Shiqian Wu; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6577-6586
Robust principal component analysis (RPCA) and its variants have gained wide app
lications in computer vision. However, these methods either involve manual adjus
tment of some parameters, or require the rank of a low-rank matrix to be known a
 prior. In this paper, an adaptive rank estimate based RPCA (ARE-RPCA) is propos
ed, which adaptively assigns weights on different singular values via rank estim
ation. More specifically, we study the characteristics of the low-rank matrix, a
nd develop an improved Gerschgorin disk theorem to estimate the rank of the low-
rank matrix accurately. Furthermore in view of the issue occurred in the Gerschg
orin disk theorem that adjustment factor need to be manually pre-defined, an ada
ptive setting method, which greatly facilitates the practical implementation of
the rank estimation, is presented. Then, the weights of singular values in the n
uclear norm are updated adaptively based on iteratively estimated rank, and the
resultant low-rank matrix is close to the target. Experimental results show that
 the proposed ARE-RPCA outperforms the state-of-the-art methods in various compl
ex scenarios.
****
Continual Adaptation of Visual Representations via Domain Randomization and Meta
-Learning
Riccardo Volpi, Diane Larlus, Gregory Rogez; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4443-4453
Most standard learning approaches lead to fragile models which are prone to drif
t when sequentially trained on samples of a different nature -- the well-known "

catastrophic forgetting" issue. In particular, when a model consecutively learns from different visual domains, it tends to forget the past domains in favor of the most recent ones. In this context, we show that one way to learn models that are inherently more robust against forgetting is domain randomization -- for vision tasks, randomizing the current domain's distribution with heavy image manipulations. Building on this result, we devise a meta-learning strategy where a regularizer explicitly penalizes any loss associated with transferring the model from the current domain to different "auxiliary" meta-domains, while also easing adaptation to them. Such meta-domains are also generated through randomized image manipulations. We empirically demonstrate in a variety of experiments -- spanning from classification to semantic segmentation -- that our approach results in models that are less prone to catastrophic forgetting when transferred to new domains.
****
DeepACG: Co-Saliency Detection via Semantic-Aware Contrast Gromov-Wasserstein Distance
Kaihua Zhang, Mingliang Dong, Bo Liu, Xiao-Tong Yuan, Qingshan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13703-13712
The objective of co-saliency detection is to segment the co-occurring salient objects in a group of images. To address this task, we introduce a new deep network architecture via semantic-aware contrast Gromov-Wasserstein distance (DeepACG). We first adopt the Gromov-Wasserstein (GW) distance to build dense hierarchical 4D correlation volumes for all pairs of image pixels within the image group. This dense correlation volumes enables the network to accurately discover the structured pair-wise pixel similarities among the common salient objects. Second, we develop a semantic-aware co-attention module (SCAM) to enhance the foreground saliency through predicted categorical information. Specifically, SCAM recognizes the semantic class of the foreground objects; and this information is then projected to the deep representations to localize the related pixels. Third, we design a contrast edge enhanced module (EEM) to capture richer context and preserve fine-grained spatial information. We validate the effectiveness of our model using three popular benchmark datasets (Cosal2015, CoSOD3k and CoCA). Extensive experiments have demonstrated the substantial practical merit of each module. Compared with the existing works, DeepACG shows significant improvements and achieves state-of-the-art performance. Code will be made available soon.
****
SurFree: A Fast Surrogate-Free Black-Box Attack
Thibault Maho, Teddy Furon, Erwan Le Merrer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10430-10439
Machine learning classifiers are critically prone to evasion attacks. Adversarial examples are slightly modified inputs that are then misclassified, while remaining perceptively close to their originals. Last couple of years have witnessed a striking decrease in the amount of queries a black box attack submits to the target classifier, in order to forge adversarials. This particularly concerns the black box score-based setup, where the attacker has access to top predicted probabilites: the amount of queries went from to millions of to less than a thousand. This paper presents SurFree, a geometrical approach that achieves a similar drastic reduction in the amount of queries in the hardest setup: black box decision-based attacks (only the top-1 label is available). We first highlight that the most recent attacks in that setup, HSJA, QEBA and GeoDA all perform costly gradient surrogate estimations. SurFree proposes to bypass these, by instead focusing on careful trials along diverse directions, guided by precise indications of geometrical properties of the classifier decision boundaries. We motivate this geometric approach before performing a head-to-head comparison with previous attacks with the amount of queries as a first class citizen. We exhibit a faster distortion decay under low query amounts (few hundreds to a thousand), while remaining competitive at higher query budgets.
****
Beyond Image to Depth: Improving Depth Prediction Using Echoes

Kranti Kumar Parida, Siddharth Srivastava, Gaurav Sharma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8268-8277

We address the problem of estimating depth with multi modal audio visual data. Inspired by the ability of animals, such as bats and dolphins, to infer distance of objects with echolocation, some recent methods have utilized echoes for depth estimation. We propose an end-to-end deep learning based pipeline utilizing RGB images, binaural echoes and estimated material properties of various objects within a scene. We argue that the relation between image, echoes and depth, for different scene elements, is greatly influenced by the properties of those elements, and a method designed to leverage this information can lead to significantly improved depth estimation from audio visual inputs. We propose a novel multi modal fusion technique, which incorporates the material properties explicitly while combining audio (echoes) and visual modalities to predict the scene depth. We show empirically, with experiments on Replica dataset, that the proposed method obtains 28% improvement in RMSE compared to the state-of-the-art audio-visual depth prediction method. To demonstrate the effectiveness of our method on larger dataset, we report competitive performance on Matterport3D, proposing to use it as a multi modal depth prediction benchmark with echoes for the first time. We also analyse the proposed method with exhaustive ablation experiments and qualitative results.

****

Rich Features for Perceptual Quality Assessment of UGC Videos

Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, Feng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13435-13444

Video quality assessment for User Generated Content (UGC) is an important topic in both industry and academia. Most existing methods only focus on one aspect of the perceptual quality assessment, such as technical quality or compression artifacts. In this paper, we create a large scale dataset to comprehensively investigate characteristics of generic UGC video quality. Besides the subjective ratings and content labels of the dataset, we also propose a DNN-based framework to thoroughly analyze importance of content, technical quality, and compression level in perceptual quality. Our model is able to provide quality scores as well as human-friendly quality indicators, to bridge the gap between low level video signals to human perceptual quality. Experimental results show that our model achieves state-of-the-art correlation with Mean Opinion Scores (MOS).

****

Sequential Graph Convolutional Network for Active Learning

Razvan Caramalau, Binod Bhattarai, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9583-9592

We propose a novel pool-based Active Learning frame-work constructed on a sequential Graph Convolution Net-work (GCN). Each image's feature from a pool of data rep-resents a node in the graph and the edges encode their similarities. With a small number of randomly sampled images as seed labelled examples, we learn the parameters of the graph to distinguish labelled vs unlabelled nodes by minimising the binary cross-entropy loss. GCN performs message-passing operations between the nodes, and hence, induces similar representations of the strongly associated nodes. We exploit these characteristics of GCN to select the unlabelled examples which are sufficiently different from la-belled ones. To this end, we utilise the graph node embed-dings and their confidence scores and adapt sampling techniques such as CoreSet and uncertainty-based methods to query the nodes. We flip the label of newly queried nodes from unlabelled to labelled, re-train the learner to optimise the downstream task and the graph to minimise its modified objective. We continue this process within a fixed budget. We evaluate our method on 6 different benchmarks: 4 real image classification, 1 depth-based hand pose estimation and 1 synthetic RGB image classification datasets. Our method outperforms several competitive baselines such as VAAL, Learning Loss, CoreSet and attains the new state-of-the-art performance on multiple applications.

****

Generative Classifiers as a Basis for Trustworthy Image Classification
Radek Mackowiak, Lynton Ardizzone, Ullrich Kothe, Carsten Rother; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2971-2981
With the maturing of deep learning systems, trustworthiness is becoming increasingly important for model assessment. We understand trustworthiness as the combination of explainability and robustness. Generative classifiers (GCs) are a promising class of models that are said to naturally accomplish these qualities. However, this has mostly been demonstrated on simple datasets such as MNIST and CIFAR in the past. In this work, we firstly develop an architecture and training scheme that allows GCs to operate on a more relevant level of complexity for practical computer vision, namely the ImageNet challenge. Secondly, we demonstrate the immense potential of GCs for trustworthy image classification. Explainability and some aspects of robustness are vastly improved compared to feed-forward models, even when the GCs are just applied naively. While not all trustworthiness problems are solved completely, we observe that GCs are a highly promising basis for further algorithms and modifications. We release our trained model for download in the hope that it serves as a starting point for other generative classification tasks, in much the same way as pretrained ResNet architectures do for discriminative classification.
****
EffiScene: Efficient Per-Pixel Rigidity Inference for Unsupervised Joint Learning of Optical Flow, Depth, Camera Pose and Motion Segmentation
Yang Jiao, Trac D. Tran, Guangming Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5538-5547
This paper addresses the challenging unsupervised scene flow estimation problem by jointly learning four low-level vision sub-tasks: optical flow F, stereo-depth D, camera pose P and motion segmentation S. Our key insight is that the rigidity of the scene shares the same inherent geometrical structure with object movements and scene depth. Hence, rigidity from S can be inferred by jointly coupling F, D and S to achieve more robust estimation. To this end, we propose a novel scene flow framework named EffiScene with efficient joint rigidity learning, going beyond the existing pipeline with independent auxiliary structures. In EffiScene, we first estimate optical flow and depth at the coarse level and then compute camera pose by Perspective-n-Points method. To jointly learn local rigidity, we design a novel Rigidity From Motion (RfM) layer with three principal components: (i) correlation extraction; (ii) boundary learning; and (iii) outlier exclusion. Final outputs are fused based on the rigid map M_R from RfM at finer levels. To efficiently train EffiScene, two new losses L_bnd and L_unc are designed to prevent trivial solutions and to regularize the flow boundary discontinuity. Extensive experiments on scene flow benchmark KITTI show that our method is effective and significantly improves the state-of-the-art approaches for all sub-tasks, i.e. optical flow (5.19 -> 4.20), depth estimation (3.78 -> 3.46), visual odometry (0.012 -> 0.011) and motion segmentation (0.57 -> 0.62).
****
Localizing Visual Sounds the Hard Way
Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16867-16876
The objective of this work is to localize sound sources that are visible in a video without using manual annotations. Our key technical contribution is to show that, by training the network to explicitly discriminate challenging image fragments, even for images that do contain the object emitting the sound, we can significantly boost the localization performance. We do so elegantly by introducing a mechanism to mine hard samples and add them to a contrastive learning formulation automatically. We show that our algorithm achieves state-of-the-art performance on the popular Flickr SoundNet dataset. Furthermore, we introduce the VGG-Sound Source (VGG-SS) benchmark, a new set of annotations for the recently-introduced VGG-Sound dataset, where the sound sources visible in each video clip are explicitly marked with bounding box annotations. This dataset is 20 times larger t

han analogous existing ones, contains 5K videos spanning over 200 categories, and, differently from Flickr SoundNet, is video-based. On VGG-SS, we also show that our algorithm achieves state-of-the-art performance against several baselines. Code and datasets can be found at http://www.robots.ox.ac.uk/ vgg/research/lvs/
****

Synthesize-It-Classifier: Learning a Generative Classifier Through Recurrent Self-Analysis
Arghya Pal, Raphael C.-W. Phan, KokSheik Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5161-5170
In this work, we show the generative capability of an image classifier network by synthesizing high-resolution, photo-realistic, and diverse images at scale. The overall methodology, called Synthesize-It-Classifier (STIC), does not require an explicit generator network to estimate the density of the data distribution and sample images from that, but instead uses the classifier's knowledge of the boundary to perform gradient ascent w.r.t. class logits and then synthesizes images using Gram Matrix Metropolis Adjusted Langevin Algorithm (GRMALA) by drawing on a blank canvas. During training, the classifier iteratively uses these synthesized images as fake samples and re-estimates the class boundary in a recurrent fashion to improve both the classification accuracy and quality of synthetic images. The STIC shows that mixing of the hard fake samples (i.e. those synthesized by the one hot class conditioning), and the soft fake samples (which are synthesized as a convex combination of classes, i.e. a mixup of classes) improves class interpolation. We demonstrate an Attentive-STIC network that shows iterative drawing of synthesized images on the ImageNet dataset that has thousands of classes. In addition, we introduce the synthesis using a class conditional score classifier (Score-STIC) instead of a normal image classifier and show improved results on several real world datasets, i.e. ImageNet, LSUN and CIFAR 10.
****

Self-Point-Flow: Self-Supervised Scene Flow Estimation From Point Clouds With Optimal Transport and Random Walk
Ruibo Li, Guosheng Lin, Lihua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15577-15586
Due to the scarcity of annotated scene flow data, self-supervised scene flow learning in point clouds has attracted increasing attention. In the self-supervised manner, establishing correspondences between two point clouds to approximate scene flow is an effective approach. Previous methods often obtain correspondences by applying point-wise matching that only takes the distance on 3D point coordinates into account, introducing two critical issues: (1) it overlooks other discriminative measures, such as color and surface normal, which often bring fruitful clues for accurate matching; and (2) it often generates sub-par performance, as the matching is operated in an unconstrained situation, where multiple points can be ended up with the same corresponding point. To address the issues, we formulate this matching task as an optimal transport problem. The output optimal assignment matrix can be utilized to guide the generation of pseudo ground truth. In this optimal transport, we design the transport cost by considering multiple descriptors and encourage one-to-one matching by mass equality constraints. Also, constructing a graph on the points, a random walk module is introduced to encourage the local consistency of the pseudo labels. Comprehensive experiments on FlyingThings3D and KITTI show that our method achieves state-of-the-art performance among self-supervised learning methods. Our self-supervised method even performs on par with some supervised learning approaches, although we do not need any ground truth flow for training.
****

Toward Joint Thing-and-Stuff Mining for Weakly Supervised Panoptic Segmentation
Yunhang Shen, Liujuan Cao, Zhiwei Chen, Feihong Lian, Baochang Zhang, Chi Su, Yongjian Wu, Feiyue Huang, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16694-16705
Panoptic segmentation aims to partition an image to object instances and semantic content for thing and stuff categories, respectively. To date, learning weakly supervised panoptic segmentation (WSPS) with only image-level labels remains un

explored. In this paper, we propose an efficient jointly thing-and-stuff mining (JTSM) framework for WSPS. To this end, we design a novel mask of interest pooling (MoIPool) to extract fixed-size pixel-accurate feature maps of arbitrary-shape segmentations. MoIPool enables a panoptic mining branch to leverage multiple instance learning (MIL) to recognize things and stuff segmentation in a unified manner. We further refine segmentation masks with parallel instance and semantic segmentation branches via self-training, which collaborates the mined masks from panoptic mining with bottom-up object evidence as pseudo-ground-truth labels to improve spatial coherence and contour localization. Experimental results demonstrate the effectiveness of JTSM on PASCAL VOC and MS COCO. As a by-product, we achieve competitive results for weakly supervised object detection and instance segmentation. This work is a first step towards tackling challenge panoptic segmentation task with only image-level labels.
****
Intelligent Carpet: Inferring 3D Human Pose From Tactile Signals
Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma, Tomas Palacios, Antonio Torralba, Wojciech Matusik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11255-11265
Daily human activities, e.g., locomotion, exercises, and resting, are heavily guided by the tactile interactions between the human and the ground. In this work, leveraging such tactile interactions, we propose a 3D human pose estimation approach using the pressure maps recorded by a tactile carpet as input. We build a low-cost, high-density, large-scale intelligent carpet, which enables the real-time recordings of human-floor tactile interactions in a seamless manner. We collect a synchronized tactile and visual dataset on various human activities. Employing a state-of-the-art camera-based pose estimation model as supervision, we design and implement a deep neural network model to infer 3D human poses using only the tactile information. Our pipeline can be further scaled up to multi-person pose estimation. We evaluate our system and demonstrate its potential applications in diverse fields.
****
Railroad Is Not a Train: Saliency As Pseudo-Pixel Supervision for Weakly Supervised Semantic Segmentation
Seungho Lee, Minhyun Lee, Jongwuk Lee, Hyunjung Shim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5495-5505
Existing studies in weakly-supervised semantic segmentation (WSSS) using image-level weak supervision have several limitations: sparse object coverage, inaccurate object boundaries, and co-occurring pixels from non-target objects. To overcome these challenges, we propose a novel framework, namely Explicit Pseudo-pixel Supervision (EPS), which learns from pixel-level feedback by combining two weak supervisions; the image-level label provides the object identity via the localization map and the saliency map from the off-the-shelf saliency detection model offers rich boundaries. We devise a joint training strategy to fully utilize the complementary relationship between both information. Our method can obtain accurate object boundaries and discard co-occurring pixels, thereby significantly improving the quality of pseudo-masks. Experimental results show that the proposed method remarkably outperforms existing methods by resolving key challenges of WSSS and achieves the new state-of-the-art performance on both PASCAL VOC 2012 and MS COCO 2014 datasets. The code is available at https://github.com/halbielee/EPS.
****
Stable View Synthesis
Gernot Riegler, Vladlen Koltun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12216-12225
We present Stable View Synthesis (SVS). Given a set of source images depicting a scene from freely distributed viewpoints, SVS synthesizes new views of the scene. The method operates on a geometric scaffold computed via structure-from-motion and multi-view stereo. Each point on this 3D scaffold is associated with view rays and corresponding feature vectors that encode the appearance of this point

in the input images. The core of SVS is view-dependent on-surface feature aggreg
ation, in which directional feature vectors at each 3D point are processed to pr
oduce a new feature vector for a ray that maps this point into the new target vi
ew. The target view is then rendered by a convolutional network from a tensor of
 features synthesized in this way for all pixels. The method is composed of diff
erentiable modules and is trained end-to-end. It supports spatially-varying view
-dependent importance weighting and feature transformation of source images at e
ach point; spatial and temporal stability due to the smooth dependence of on-sur
face feature aggregation on the target view; and synthesis of view-dependent eff
ects such as specular reflection. Experimental results demonstrate that SVS outp
erforms state-of-the-art view synthesis methods both quantitatively and qualitat
ively on three diverse real-world datasets, achieving unprecedented levels of re
alism in free-viewpoint video of challenging large-scale scenes. Code is availab
le at https://github.com/intel-isl/StableViewSynthesis
****
Deep Two-View Structure-From-Motion Revisited
Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai S
molyanskiy, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2021, pp. 8953-8962
Two-view structure-from-motion (SfM) is the cornerstone of 3D reconstruction and
 visual SLAM. Existing deep learning-based approaches formulate the problem in w
ays that are fundamentally ill-posed, relying on training data to overcome the i
nherent difficulties. In contrast, we propose a return to the basics. We revisit
 the problem of deep two-view SfM by leveraging the well-posedness of the classi
c pipeline. Our method consists of 1) an optical flow estimation network that pr
edicts dense correspondences between two frames; 2) a normalized pose estimation
 module that computes relative camera poses from the 2D optical flow corresponde
nces, and 3) a scale-invariant depth estimation network that leverages epipolar
geometry to reduce the search space, refine the dense correspondences, and estim
ate relative depth maps. Extensive experiments show that our method outperforms
all state-of-the-art two-view SfM methods by a clear margin on KITTI depth, KITT
I VO, MVS, Scenes11, and SUN3D datasets in both relative pose estimation and dep
th estimation.
****
Rethinking Style Transfer: From Pixels to Parameterized Brushstrokes
Dmytro Kotovenko, Matthias Wright, Arthur Heimbrecht, Bjorn Ommer; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 12196-12205
There have been many successful implementations of neural style transfer in rece
nt years. In most of these works, the stylization process is confined to the pix
el domain. However, we argue that this representation is unnatural because paint
ings usually consist of brushstrokes rather than pixels. We propose a method to
stylize images by optimizing parameterized brushstrokes instead of pixels and fu
rther introduce a simple differentiable rendering mechanism. Our approach signif
icantly improves visual quality and enables additional control over the stylizat
ion process such as controlling the flow of brushstrokes through user input. We
provide qualitative and quantitative evaluations that show the efficacy of the p
roposed parameterized representation.
****
Cluster, Split, Fuse, and Update: Meta-Learning for Open Compound Domain Adaptiv
e Semantic Segmentation
Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengx
in Dai, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2021, pp. 8344-8354
Open compound domain adaptation (OCDA) is a domain adaptation setting, where tar
get domain is modeled as a compound of multiple unknown homogeneous domains, whi
ch brings the advantage of improved generalization to unseen domains. In this wo
rk, we propose a principled meta-learning based approach to OCDA for semantic se
gmentation, MOCDA, by modeling the unlabeled target domain continuously. Our app
roach consists of four key steps. First, we cluster target domain into multiple

sub-target domains by image styles, extracted in an unsupervised manner. Then, d
ifferent sub-target domains are split into independent branches, for which batch
 normalization parameters are learnt to treat them independently. A meta-learner
 is thereafter deployed to learn to fuse sub-target domain-specific predictions,
 conditioned upon the style code. Meanwhile, we learn to online update the model
 by model-agnostic meta-learning (MAML) algorithm, thus to further improve gener
alization. We validate the benefits of our approach by extensive experiments on
synthetic-to-real knowledge transfer benchmark, where we achieve the state-of-th
e-art performance in both compound and open domains.
****
Beyond Short Clips: End-to-End Video-Level Learning With Collaborative Memories
Xitong Yang, Haoqi Fan, Lorenzo Torresani, Larry S. Davis, Heng Wang; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2021, pp. 7567-7576
The standard way of training video models entails sampling at each iteration a s
ingle clip from a video and optimizing the clip prediction with respect to the v
ideo-level label. We argue that a single clip may not have enough temporal cover
age to exhibit the label to recognize, since video datasets are often weakly lab
eled with categorical information but without dense temporal annotations. Furthe
rmore, optimizing the model over brief clips impedes its ability to learn long-t
erm temporal dependencies. To overcome these limitations, we introduce a collabo
rative memory mechanism that encodes information across multiple sampled clips o
f a video at each training iteration. This enables the learning of long-range de
pendencies beyond a single clip. We explore different design choices for the col
laborative memory to ease the optimization difficulties. Our proposed framework
is end-to-end trainable and significantly improves the accuracy of video classif
ication at a negligible computational overhead. Through extensive experiments, w
e demonstrate that our framework generalizes to different video architectures an
d tasks, outperforming the state of the art on both action recognition (e.g., Ki
netics-400 & 700, Charades, Something-Something-V1) and action detection (e.g.,
AVA v2.1 & v2.2).
****
PointDSC: Robust Point Cloud Registration Using Deep Spatial Consistency
Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, Chiew
-Lan Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2021, pp. 15859-15869
Removing outlier correspondences is one of the critical steps for successful fea
ture-based point cloud registration. Despite the increasing popularity of introd
ucing deep learning methods in this field, spatial consistency, which is essenti
ally established by a Euclidean transformation between point clouds, has receive
d almost no individual attention in existing learning frameworks. In this paper,
 we present PointDSC, a novel deep neural network that explicitly incorporates s
patial consistency for pruning outlier correspondences. First, we propose a nonl
ocal feature aggregation module, weighted by both feature and spatial coherence,
 for feature embedding of the input correspondences. Second, we formulate a diff
erentiable spectral matching module, supervised by pairwise spatial compatibilit
y, to estimate the inlier confidence of each correspondence from the embedded fe
atures. With modest computation cost, our method outperforms the state-of-the-ar
t hand-crafted and learning-based outlier rejection approaches on several real-w
orld datasets by a significant margin. We also show its wide applicability by co
mbining PointDSC with different 3D local descriptors.
****
Task Programming: Learning Data Efficient Behavior Representations
Jennifer J. Sun, Ann Kennedy, Eric Zhan, David J. Anderson, Yisong Yue, Pietro P
erona; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2021, pp. 2876-2885
Specialized domain knowledge is often necessary to accurately annotate training
sets for in-depth analysis, but can be burdensome and time-consuming to acquire
from domain experts. This issue arises prominently in automated behavior analysi
s, in which agent movements or actions of interest are detected from video track

ing data. To reduce annotation effort, we present TREBA: a method to learn annotation-sample efficient trajectory embedding for behavior analysis, based on multi-task self-supervised learning. The tasks in our method can be efficiently engineered by domain experts through a process we call "task programming", which uses programs to explicitly encode structured knowledge from domain experts. Total domain expert effort can be reduced by exchanging data annotation time for the construction of a small number of programmed tasks. We evaluate this trade-off using data from behavioral neuroscience, in which specialized domain knowledge is used to identify behaviors. We present experimental results in three datasets across two domains: mice and fruit flies. Using embeddings from TREBA, we reduce annotation burden by up to a factor of 10 without compromising accuracy compared to state-of-the-art features. Our results thus suggest that task programming and self-supervision can be an effective way to reduce annotation effort for domain experts.
****
ACRE: Abstract Causal REasoning Beyond Covariation
Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, Yixin Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10643-10653
Causal induction, i.e., identifying unobservable mechanisms that lead to the observable relations among variables, has played a pivotal role in modern scientific discovery, especially in scenarios with only sparse and limited data. Humans, even young toddlers, can induce causal relationships surprisingly well in various settings despite its notorious difficulty. However, in contrast to the commonplace trait of human cognition is the lack of a diagnostic benchmark to measure causal induction for modern Artificial Intelligence (AI) systems. Therefore, in this work, we introduce the Abstract Causal REasoning (ACRE) dataset for systematic evaluation of current vision systems in causal induction. Motivated by the stream of research on causal discovery in Blicket experiments, we query a visual reasoning system with the following four types of questions in either an independent scenario or an interventional scenario: direct, indirect, screening-off, and backward-blocking, intentionally going beyond the simple strategy of inducing causal relationships by covariation. By analyzing visual reasoning architectures on this testbed, we notice that pure neural models tend towards an associative strategy under their chance-level performance, whereas neuro-symbolic combinations struggle in backward-blocking reasoning. These deficiencies call for future research in models with a more comprehensive capability of causal induction.
****
DeepLM: Large-Scale Nonlinear Least Squares on Deep Learning Frameworks Using Stochastic Domain Decomposition
Jingwei Huang, Shan Huang, Mingwei Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10308-10317
We propose a novel approach for large-scale nonlinear least squares problems based on deep learning frameworks. Nonlinear least squares are commonly solved with the Levenberg-Marquardt (LM) algorithm for fast convergence. We implement a general and efficient LM solver on a deep learning framework by designing a new backward jacobian network to enable automatic sparse jacobian matrix computation. Furthermore, we introduce a stochastic domain decomposition approach that enables batched optimization and preserves convergence for large problems. We evaluate our method by solving bundle adjustment as a fundamental problem. Experiments show that our optimizer significantly outperforms the state-of-the-art solutions and existing deep learning solvers considering quality, efficiency, and memory. Our stochastic domain decomposition enables distributed optimization, consumes little memory and time, and achieves similar quality compared to a global solver. As a result, our solver effectively solves nonlinear least squares on an extremely large scale. We will make the code publicly available on publication.
****
TDN: Temporal Difference Networks for Efficient Action Recognition
Limin Wang, Zhan Tong, Bin Ji, Gangshan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1895-1904

Temporal modeling still remains challenging for action recognition in videos. To
 mitigate this issue, this paper presents a new video architecture, termed as Te
mporal Difference Network (TDN), with a focus on capturing multi-scale temporal
information for efficient action recognition. The core of our TDN is to devise a
n efficient temporal module (TDM) by explicitly leveraging a temporal difference
 operator, and systematically assess its effect on short-term and long-term moti
on modeling. To fully capture temporal information over the entire video, our TD
N is established with a two-level difference modeling paradigm. Specifically, fo
r local motion modeling, temporal difference over consecutive frames is used to
supply 2D CNNs with finer motion pattern, while for global motion modeling, temp
oral difference across segments is incorporated to capture long-range structure
for motion feature excitation. TDN provides a simple and principled temporal mod
eling framework and could be instantiated with the existing CNNs at a small extr
a computational cost. Our TDN presents a new state of the art on the Something-S
omething V1 & V2 datasets and is on par with the best performance on the Kinetic
s-400 dataset. In addition, we conduct in-depth ablation studies and plot the vi
sualization results of our TDN, hopefully providing insightful analysis on tempo
ral difference modeling. We release the code at https://github.com/MCG-NJU/TDN.
****
LiBRe: A Practical Bayesian Approach to Adversarial Detection
Zhijie Deng, Xiao Yang, Shizhen Xu, Hang Su, Jun Zhu; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 972-98
2
Despite their appealing flexibility, deep neural networks (DNNs) are vulnerable
against adversarial examples. Various adversarial defense strategies have been p
roposed to resolve this problem, but they typically demonstrate restricted pract
icability owing to unsurmountable compromise on universality, effectiveness, or
efficiency. In this work, we propose a more practical approach, Lightweight Baye
sian Refinement (LiBRe), in the spirit of leveraging Bayesian neural networks (B
NNs) for adversarial detection. Empowered by the task and attack agnostic modeli
ng under Bayes principle, LiBRe can endow a variety of pre-trained task-dependen
t DNNs with the ability of defending heterogeneous adversarial attacks at a low
cost. We develop and integrate advanced learning techniques to make LiBRe approp
riate for adversarial detection. Concretely, we build the few-layer deep ensembl
e variational and adopt the pre-training & fine-tuning workflow to boost the eff
ectiveness and efficiency of LiBRe. We further provide a novel insight to realis
e adversarial detection-oriented uncertainty quantification without inefficientl
y crafting adversarial examples during training. Extensive empirical studies cov
ering a wide range of scenarios verify the practicability of LiBRe. We also cond
uct thorough ablation studies to evidence the superiority of our modeling and le
arning strategies.
****
ArtCoder: An End-to-End Method for Generating Scanning-Robust Stylized QR Codes
Hao Su, Jianwei Niu, Xuefeng Liu, Qingfeng Li, Ji Wan, Mingliang Xu, Tao Ren; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2021, pp. 2277-2286
Quick Response (QR) code is one of the most worldwide used two-dimensional codes
. Traditional QR codes appear as random collections of black-and-white modules t
hat lack visual semantics and aesthetic elements, which inspires the recent work
s to beautify the appearances of QR codes. However, these works typically beatif
y QR codes in a single style due to the fixed generation algorithms, which is im
provable in personalization and diversification. In this paper, combining the Ne
ural Style Transfer technique, we propose a novel end-to-end network ACN (ArtCod
e-Net) to generate the stylized QR codes that are personalized, diverse, attract
ive, and scanning-robust. To address the challenge that preserving the scanning-
robustness after giving such codes style elements, we further propose the Sampli
ng-Simulation layer, the module-based code loss, and a competition mechanism to
improve the performances of ACN. The experimental results show that our stylized
 QR codes have high-quality in both the visual effect and the scanning-robustnes
s, and they are able to support the real-world application.

****
Self-Supervised Pillar Motion Learning for Autonomous Driving
Chenxu Luo, Xiaodong Yang, Alan Yuille; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3183-3192
Autonomous driving can benefit from motion behavior comprehension when interacti
ng with diverse traffic participants in highly dynamic environments. Recently, t
here has been a growing interest in estimating class-agnostic motion directly fr
om point clouds. Current motion estimation methods usually require vast amount o
f annotated training data from self-driving scenes. However, manually labeling p
oint clouds is notoriously difficult, error-prone and time-consuming. In this pa
per, we seek to answer the research question of whether the abundant unlabeled d
ata collections can be utilized for accurate and efficient motion learning. To t
his end, we propose a learning framework that leverages free supervisory signals
 from point clouds and paired camera images to estimate motion purely via self-s
upervision. Our model involves a point cloud based structural consistency augmen
ted with probabilistic motion masking as well as a cross-sensor motion regulariz
ation to realize the desired self-supervision. Experiments reveal that our appro
ach performs competitively to supervised methods, and achieves the state-of-the-
art result when combining our self-supervised model with supervised fine-tuning.
****
Quantum Permutation Synchronization
Tolga Birdal, Vladislav Golyanik, Christian Theobalt, Leonidas J. Guibas; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2021, pp. 13122-13133
We present QuantumSync, the first quantum algorithm for solving a synchronizatio
n problem in the context of computer vision. In particular, we focus on permutat
ion synchronization which involves solving a non-convex optimization problem in
discrete variables. We start by formulating synchronization into a quadratic unc
onstrained binary optimization problem (QUBO). While such formulation respects t
he binary nature of the problem, ensuring that the result is a set of permutatio
ns requires extra care. Hence, we: (i) show how to insert permutation constraint
s into a QUBO problem and (ii) solve the constrained QUBO problem on the current
 generation of the adiabatic quantum computers D-Wave. Thanks to the quantum ann
ealing, we guarantee global optimality with high probability while sampling the
energy landscape to yield confidence estimates. Our proof-of-concepts realizatio
n on the adiabatic D-Wave computer demonstrates that quantum machines offer a pr
omising way to solve the prevalent yet difficult synchronization problems.
****
QAIR: Practical Query-Efficient Black-Box Attacks for Image Retrieval
Xiaodan Li, Jinfeng Li, Yuefeng Chen, Shaokai Ye, Yuan He, Shuhui Wang, Hang Su,
 Hui Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2021, pp. 3330-3339
We study the query-based attack against image retrieval to evaluate its robustne
ss against adversarial examples under the black-box setting, where the adversary
 only has query access to the top-k ranked unlabeled images from the database. C
ompared with query attacks in image classification, which produce adversaries ac
cording to the returned labels or confidence score, the challenge becomes even m
ore prominent due to the difficulty in quantifying the attack effectiveness on t
he partial retrieved list. In this paper, we make the first attempt in Query-bas
ed Attack against Image Retrieval (QAIR), to completely subvert the top-k retrie
val results. Specifically, a new relevance-based loss is designed to quantify th
e attack effects by measuring the set similarity on the top-k retrieval results
before and after attacks and guide the gradient optimization. To further boost t
he attack efficiency, a recursive model stealing method is proposed to acquire t
ransferable priors on the target model and generate the prior-guided gradients.
Comprehensive experiments show that the proposed attack achieves a high attack s
uccess rate with few queries against the image retrieval systems under the black
-box setting. The attack evaluations on real-world visual search engine show tha
t it successfully deceives a commercial system such as Bing Visual Search with 9
8% attack success rate by only 33 queries on average.

****
MagFace: A Universal Representation for Face Recognition and Quality Assessment
Qiang Meng, Shichao Zhao, Zhida Huang, Feng Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14225-14234
The performance of face recognition system degrades when the variability of the acquired faces increases. Prior work alleviates this issue by either monitoring the face quality in pre-processing or predicting the data uncertainty along with the face feature. This paper proposes MagFace, a category of losses that learn a universal feature embedding whose magnitude before normalization can measure with the quality of the given face. Under the new loss, it can be proven that the magnitude of the feature embedding monotonically increases if the subject is more likely to be recognized. In addition, MagFace introduces an adaptive mechanism to learn a well-structured within-class feature distributions by pushing easy samples to class centers while pushing hard samples away. This prevents models from overfitting on noisy low-quality samples and improves face recognition in the wild. Extensive experiments conducted on face recognition, quality assessments as well as clustering have demonstrated the effectiveness of MagFace over state-of-the-arts. The code is available at https://github.com/IrvingMeng/MagFace.
****
Wasserstein Barycenter for Multi-Source Domain Adaptation
Eduardo Fernandes Montesuma, Fred Maurice Ngole Mboula; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16785-16793
Multi-source domain adaptation is a key technique that allows a model to be trained on data coming from various probability distribution. To overcome the challenges posed by this learning scenario, we propose a method for constructing an intermediate domain between sources and target domain, the Wasserstein Barycenter Transport (WBT). This method relies on the barycenter on Wasserstein spaces for aggregating the source probability distributions. Once the sources have been aggregated, they are transported to the target domain using standard Optimal Transport for Domain Adaptation framework. Additionally, we revisit previous single-source domain adaptation tasks in the context of multi-source scenario. In particular, we apply our algorithm to object and face recognition datasets. Moreover, to diversify the range of applications, we also examine the tasks of music genre recognition and music-speech discrimination. The experiments show that our method has similar performance with the existing state-of-the-art.
****
Unsupervised Hyperbolic Metric Learning
Jiexi Yan, Lei Luo, Cheng Deng, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12465-12474
Learning feature embedding directly from images without any human supervision is a very challenging and essential task in the field of computer vision and machine learning. Following the paradigm in supervised manner, most existing unsupervised metric learning approaches mainly focus on binary similarity in Euclidean space. However, these methods cannot achieve promising performance in many practical applications, where the manual information is lacking and data exhibits non-Euclidean latent anatomy. To address this limitation, we propose an Unsupervised Hyperbolic Metric Learning method with Hierarchical Similarity. It considers the natural hierarchies of data by taking advantage of Hyperbolic metric learning and hierarchical clustering, which can effectively excavate richer similarity information beyond binary in modeling. More importantly, we design a new loss function to capture the hierarchical similarity among samples to enhance the stability of the proposed method. Extensive experimental results on benchmark datasets demonstrate that our method achieves state-of-the-art performance compared with current unsupervised deep metric learning approaches.
****
Improving Sign Language Translation With Monolingual Data by Sign Back-Translation
Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, Houqiang Li; Proceedings of the IE

EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1316-1325
Despite existing pioneering works on sign language translation (SLT), there is a non-trivial obstacle, i.e., the limited quantity of parallel sign-text data. To tackle this parallel data bottleneck, we propose a sign back-translation (SignBT) approach, which incorporates massive spoken language texts into SLT training. With a text-to-gloss translation model, we first back-translate the monolingual text to its gloss sequence. Then, the paired sign sequence is generated by splicing pieces from an estimated gloss-to-sign bank at the feature level. Finally, the synthetic parallel data serves as a strong supplement for the end-to-end training of the encoder-decoder SLT framework. To promote the SLT research, we further contribute CSL-Daily, a large-scale continuous SLT dataset. It provides both spoken language translations and gloss-level annotations. The topic revolves around people's daily lives (e.g., travel, shopping, medical care), the most likely SLT application scenario. Extensive experimental results and analysis of SLT methods are reported on CSL-Daily. With the proposed sign back-translation method, we obtain a substantial improvement over previous state-of-the-art SLT methods.
****
Background Splitting: Finding Rare Classes in a Sea of Background
Ravi Teja Mullapudi, Fait Poms, William R. Mark, Deva Ramanan, Kayvon Fatahalian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8043-8052
We focus on the problem of training deep image classification models for a small number of extremely rare categories. In this common, real-world scenario, almost all images belong to the background category in the dataset. We find that state-of-the-art approaches for training on imbalanced datasets do not produce accurate deep models in this regime. Our solution is to split the large, visually diverse background into many smaller, visually similar categories during training. We implement this idea by extending an image classification model with an additional auxiliary loss that learns to mimic the predictions of a pre-existing classification model on the training set. The auxiliary loss requires no additional human labels and regularizes feature learning in the shared network trunk by forcing the model to discriminate between auxiliary categories for all training set examples, including those belonging to the monolithic background of the main rare category classification task. To evaluate our method we contribute modified versions of the iNaturalist and Places365 datasets where only a small subset of rare category labels are available during training (all other images are labeled as background). By jointly learning to recognize both the selected rare categories and auxiliary categories, our approach yields models that perform 8.3 mAP points higher than state-of-the-art imbalanced learning baselines when 98.30% of the data is background, and up to 42.3 mAP points higher than fine-tuning baselines when 99.98% of the data is background.
****
Adaptive Convolutions for Structure-Aware Style Transfer
Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, Derek Bradley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7972-7981
Style transfer between images is an artistic application of CNNs, where the 'style' of one image is transferred onto another image while preserving the latter's content. The state of the art in neural style transfer is based on Adaptive Instance Normalization (AdaIN), a technique that transfers the statistical properties of style features to a content image, and can transfer a large number of styles in real time. However, AdaIN is a global operation; thus local geometric structures in the style image are often ignored during the transfer. We propose Adaptive Convolutions (AdaConv), a generic extension of AdaIN, to allow for the simultaneous transfer of both statistical and structural styles in real time. Apart from style transfer, our method can also be readily extended to style-based image generation, and other tasks where AdaIN has already been adopted.
****

Few-Shot Incremental Learning With Continually Evolved Classifiers
Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, Yinghui Xu; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 12455-12464
Few-shot class-incremental learning (FSCIL) aims to design machine learning algo
rithms that can continually learn new concepts from a few data points, without f
orgetting knowledge of old classes. The difficulty lies in that limited data fro
m new classes not only lead to significant overfitting issues but also exacerbat
e the notorious catastrophic forgetting problems. Moreover, as training data com
e in sequence in FSCIL, the learned classifier can only provide discriminative i
nformation in individual sessions, while FSCIL requires all classes to be involv
ed for evaluation. In this paper, we address the FSCIL problem from two aspects.
 First, we adopt a simple but effective decoupled learning strategy of represent
ations and classifiers that only the classifiers are updated in each incremental
 session, which avoids knowledge forgetting in the representations. By doing so,
 we demonstrate that a pre-trained backbone plus a non-parametric class mean cla
ssifier can beat state-of-the-art methods. Second, to make the classifiers learn
ed on individual sessions applicable to all classes, we propose a Continually Ev
olved Classifier (CEC) that employs a graph model to propagate context informati
on between classifiers for adaptation. To enable the learning of CEC, we design
a pseudo incremental learning paradigm that episodically constructs a pseudo inc
remental learning task to optimize the graph parameters by sampling data from th
e base dataset. Experiments on three popular benchmark datasets, including CIFAR
100, miniImageNet, and Caltech-USCD Birds-200-2011 (CUB200), show that our metho
d significantly outperforms the baselines and sets new state-of-the-art results
with remarkable advantages.
****
NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions
Junbin Xiao, Xindi Shang, Angela Yao, Tat-Seng Chua; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9777-97
86
We introduce NExT-QA, a rigorously designed video question answering (VideoQA) b
enchmark to advance video understanding from describing to explaining the tempor
al actions. Based on the dataset, we set up multi-choice and open-ended QA tasks
 targeting at causal action reasoning, temporal action reasoning and common scen
e comprehension. Through extensive analysis of baselines and established VideoQA
 techniques, we find that top-performing methods excel at shallow scene descript
ions but are weak in causal and temporal action reasoning. Furthermore, the mode
ls that are effective on multi-choice QA, when adapted to open-ended QA, still s
truggle in generalizing the answers. This raises doubt on the ability of these m
odels to reason and highlights possibilities for improvement. With detailed resu
lts for different question types and heuristic observations for future works, we
 hope NExT-QA will guide the next generation of VQA research to go beyond superf
icial description towards a deeper understanding of videos.
****
LayoutGMN: Neural Graph Matching for Structural Layout Similarity
Akshay Gadi Patil, Manyi Li, Matthew Fisher, Manolis Savva, Hao Zhang; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2021, pp. 11048-11057
We present a deep neural network to predict structural similarity between 2D lay
outs by leveraging Graph Matching Networks (GMN). Our network, coined LayoutGMN,
 learns the layout metric via neural graph matching, using an attention-based GM
N designed under a triplet network setting. To train our network, we utilize wea
k labels obtained by pixel-wise Intersection-over-Union (IoUs) to define the tri
plet loss. Importantly, LayoutGMN is built with a structural bias which can effe
ctively compensate for the lack of structure awareness in IoUs. We demonstrate t
his on two prominent forms of layouts, viz., floorplans and UI designs, via retr
ieval experiments on large-scale datasets. In particular, retrieval results by o
ur network better match human judgement of structural layout similarity compared
 to both IoUs and other baselines including a state-of-the-art method based on g

raph neural networks and image convolution. In addition, LayoutGMN is the first deep model to offer both metric learning of structural layout similarity and structural matching between layout elements.
****
TransNAS-Bench-101: Improving Transferability and Generalizability of Cross-Task Neural Architecture Search
Yawen Duan, Xin Chen, Hang Xu, Zewei Chen, Xiaodan Liang, Tong Zhang, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5251-5260
Recent breakthroughs of Neural Architecture Search (NAS) extend the field's research scope towards a broader range of vision tasks and more diversified search spaces. While existing NAS methods mostly design architectures on a single task, algorithms that look beyond single-task search are surging to pursue a more efficient and universal solution across various tasks. Many of them leverage transfer learning and seek to preserve, reuse, and refine network design knowledge to achieve higher efficiency in future tasks. However, the enormous computational cost and experiment complexity of cross-task NAS are imposing barriers for valuable research in this direction. Existing NAS benchmarks all focus on one type of vision task, i.e., classification. In this work, we propose TransNAS-Bench-101, a benchmark dataset containing network performance across seven tasks, covering classification, regression, pixel-level prediction, and self-supervised tasks. This diversity provides opportunities to transfer NAS methods among tasks and allows for more complex transfer schemes to evolve. We explore two fundamentally different types of search space: cell-level search space and macro-level search space. With 7,352 backbones evaluated on seven tasks, 51,464 trained models with detailed training information are provided. With TransNAS-Bench-101, we hope to encourage the advent of exceptional NAS algorithms that raise cross-task search efficiency and generalizability to the next level. Our dataset and code will be available at Mindspore and VEGA.
****
ArtEmis: Affective Language for Visual Art
Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11569-11579
We present a novel large-scale dataset and accompanying machine learning models aimed at providing a detailed understanding of the interplay between visual content, its emotional effect, and explanations for the latter in language. In contrast to most existing annotation datasets in computer vision, we focus on the affective experience triggered by visual artworks and ask the annotators to indicate the dominant emotion they feel for a given image and, crucially, to also provide a grounded verbal explanation for their emotion choice. As we demonstrate below, this leads to a rich set of signals for both the objective content and the affective impact of an image, creating associations with abstract concepts (e.g., "freedom" or "love"), or references that go beyond what is directly visible, including visual similes and metaphors, or subjective references to personal experiences. We focus on visual art (e.g., paintings, artistic photographs) as it is a prime example of imagery created to elicit emotional responses from its viewers. Our dataset, termed ArtEmis, contains 455K emotion attributions and explanations from humans, on 80K artworks from WikiArt. Building on this data, we train and demonstrate a series of captioning systems capable of expressing and explaining emotions from visual stimuli. Remarkably, the captions produced by these systems often succeed in reflecting the semantic and abstract content of the image, going well beyond systems trained on existing datasets.
****
Sketch, Ground, and Refine: Top-Down Dense Video Captioning
Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 234-243
The dense video captioning task aims to detect and describe a sequence of events in a video for detailed and coherent storytelling. Previous works mainly adopt a "detect-then-describe" framework, which firstly detects event proposals in the

video and then generates descriptions for the detected events. However, the definitions of events are diverse which could be as simple as a single action or as complex as a set of events, depending on different semantic contexts. Therefore, directly detecting events based on video information is ill-defined and hurts the coherency and accuracy of generated dense captions. In this work, we reverse the predominant "detect-then-describe" fashion, proposing a top-down way to first generate paragraphs from a global view and then ground each event description to a video segment for detailed refinement. It is formulated as a Sketch, Ground, and Refine process (SGR). The sketch stage first generates a coarse-grained multi-sentence paragraph to describe the whole video, where each sentence is treated as an event and gets localised in the grounding stage. In the refining stage, we improve captioning quality via refinement-enhanced training and dual-path cross attention on both coarse-grained event captions and aligned event segments. The updated event caption can further adjust its segment boundaries. Our SGR model outperforms state-of-the-art methods on ActivityNet Captioning benchmark under traditional and story-oriented dense caption evaluations. Code will be released at github.com/bearcatt/SGR.
****
Learning Normal Dynamics in Videos With Meta Prototype Network
Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15425-15434
Frame reconstruction (current or future frames) based on Auto-Encoder (AE) is a popular method for video anomaly detection. With models trained on the normal data, the reconstruction errors of anomalous scenes are usually much larger than those of normal ones. Previous methods introduced the memory bank into AE, for encoding diverse normal patterns across the training videos. However, they are memory-consuming and cannot cope with unseen new scenarios in the training data. In this work, we propose a dynamic prototype unit (DPU) to encode the normal dynamics as prototypes in real time, free from extra memory cost. In addition, we introduce meta-learning to our DPU to form a novel few-shot normalcy learner, namely Meta-Prototype Unit (MPU). It enables the fast adaption capability on new scenes by only consuming a few iterations of update. Extensive experiments are conducted on various benchmarks. The superior performance over the state-of-the-art demonstrates the effectiveness of our method.
****
Graph-Based High-Order Relation Discovery for Fine-Grained Recognition
Yifan Zhao, Ke Yan, Feiyue Huang, Jia Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15079-15088
Fine-grained object recognition aims to learn effective features that can identify the subtle differences between visually similar objects. Most of the existing works tend to amplify discriminative part regions with attention mechanisms. Besides its unstable performance under complex backgrounds, the intrinsic interrelationship between different semantic features is less explored. Toward this end, we propose an effective graph-based relation discovery approach to build a contextual understanding of high-order relationships. In our approach, a high-dimensional feature bank is first formed and jointly regularized with semantic- and positional-aware high-order constraints, endowing rich attributes to feature representations. Second, to overcome the high-dimension curse, we propose a graph-based semantic grouping strategy to embed this high-order tensor bank into a low-dimensional space. Meanwhile, a group-wise learning strategy is proposed to regularize the features focusing on the cluster embedding center. With the collaborative learning of three modules, our module is able to grasp the stronger contextual details of fine-grained objects. Experimental evidence demonstrates our approach achieves new state-of-the-art on 4 widely-used fine-grained object recognition benchmarks.
****
Normal Integration via Inverse Plane Fitting With Minimum Point-to-Plane Distance
Xu Cao, Boxin Shi, Fumio Okura, Yasuyuki Matsushita; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2382-2391
This paper presents a surface normal integration method that solves an inverse problem of local plane fitting. Surface reconstruction from normal maps is essential in photometric shape reconstruction. To this end, we formulate normal integration in the camera coordinates and jointly solve for 3D point positions and local plane displacements. Unlike existing methods that consider the vertical distances between 3D points, we minimize the sum of squared point-to-plane distances. Our method can deal with both orthographic or perspective normal maps with arbitrary boundaries. Compared to existing normal integration methods, our method avoids the checkerboard artifact and performs more robustly against natural boundaries, sharp features, and outliers. We further provide a geometric analysis of the source of artifacts that appear in previous methods based on our plane fitting formulation. Experimental results on analytically computed, synthetic, and real-world surfaces show that our method yields accurate and stable reconstruction for both orthographic and perspective normal maps.
****
NPAS: A Compiler-Aware Framework of Unified Network Pruning and Architecture Search for Beyond Real-Time Mobile Acceleration
Zhengang Li, Geng Yuan, Wei Niu, Pu Zhao, Yanyu Li, Yuxuan Cai, Xuan Shen, Zheng Zhan, Zhenglun Kong, Qing Jin, Zhiyu Chen, Sijia Liu, Kaiyuan Yang, Bin Ren, Yanzhi Wang, Xue Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14255-14266
With the increasing demand to efficiently deploy DNNs on mobile edge devices, it becomes much more important to reduce unnecessary computation and increase the execution speed. Prior methods towards this goal, including model compression and network architecture search (NAS), are largely performed independently, and do not fully consider compiler-level optimizations which is a must-do for mobile acceleration. In this work, we first propose (i) a general category of fine-grained structured pruning applicable to various DNN layers, and (ii) a comprehensive, compiler automatic code generation framework supporting different DNNs and different pruning schemes, which bridge the gap of model compression and NAS. We further propose NPAS, a compiler-aware unified network pruning and architecture search. To deal with large search space, we propose a meta-modeling procedure based on reinforcement learning with fast evaluation and Bayesian optimization, ensuring the total number of training epochs comparable with representative NAS frameworks. Our framework achieves 6.7ms, 5.9ms, and 3.9ms ImageNet inference times with 78.2%, 75% (MobileNet-V3 level), and 71% (MobileNet-V2 level) Top-1 accuracy respectively on an off-the-shelf mobile phone, consistently outperforming prior work.
****
Spatial Feature Calibration and Temporal Fusion for Effective One-Stage Video Instance Segmentation
Minghan Li, Shuai Li, Lida Li, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11215-11224
Modern one-stage video instance segmentation networks suffer from two limitations. First, convolutional features are neither aligned with anchor boxes nor with ground-truth bounding boxes, reducing the mask sensitivity to spatial location. Second, a video is directly divided into individual frames for frame-level instance segmentation, ignoring the temporal correlation between adjacent frames. To address these issues, we propose a simple yet effective one-stage video instance segmentation framework by spatial calibration and temporal fusion, namely STMask. To ensure spatial feature calibration with ground-truth bounding boxes, we first predict regressed bounding boxes around ground-truth bounding boxes, and extract features from them for frame-level instance segmentation. To further explore temporal correlation among video frames, we aggregate a temporal fusion module to infer instance masks from each frame to its adjacent frames, which helps our framework to handle challenging videos such as motion blur, partial occlusion and unusual object-to-camera poses. Experiments on the YouTube-VIS valid set show that the proposed STMask with ResNet-50/-101 backbone obtains 33.5 % / 36.8 % m

ask AP, while achieving 28.6 / 23.4 FPS on video instance segmentation. The code is released online https://github.com/MinghanLi/STMask.
****
Learning Asynchronous and Sparse Human-Object Interaction in Videos
Romero Morais, Vuong Le, Svetha Venkatesh, Truyen Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16041-16050
Human activities can be learned from video. With effective modeling it is possible to discover not only the action labels but also the temporal structure of the activities, such as the progression of the sub-activities. Automatically recognizing such structure from raw video signal is a new capability that promises authentic modeling and successful recognition of human-object interactions. Toward this goal, we introduce Asynchronous-Sparse Interaction Graph Networks (ASSIGN), a recurrent graph network that is able to automatically detect the structure of interaction events associated with entities in a video scene. ASSIGN pioneers learning of autonomous behavior of video entities including their dynamic structure and their interaction with the coexisting neighbors. Entities' lives in our model are asynchronous to those of others therefore more flexible in adapting to complex scenarios. Their interactions are sparse in time hence more faithful to the true underlying nature and more robust in inference and learning. ASSIGN is tested on human-object interaction recognition and shows superior performance in segmenting and labeling of human sub-activities and object affordances from raw videos. The native ability of ASSIGN in discovering temporal structure also eliminates the dependence on external segmentation that was previously mandatory for this task.
****
Single Image Reflection Removal With Absorption Effect
Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, Alex C. Kot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13395-13404
In this paper, we consider the absorption effect for the problem of single image reflection removal. We show that the absorption effect can be numerically approximated by the average of refractive amplitude coefficient map. We then reformulate the image formation model and propose a two-step solution that explicitly takes the absorption effect into account. The first step estimates the absorption effect from a reflection-contaminated image, while the second step recovers the transmission image by taking a reflection-contaminated image and the estimated absorption effect as the input. Experimental results on four public datasets show that our two-step solution not only successfully removes reflection artifact, but also faithfully restores the intensity distortion caused by the absorption effect. Our ablation studies further demonstrate that our method achieves superior performance on the recovery of overall intensity and has good model generalization capacity. The code is available at https://github.com/q-zh/absorption.
****
One-Shot Neural Ensemble Architecture Search by Diversity-Guided Search Space Shrinking
Minghao Chen, Jianlong Fu, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16530-16539
Despite remarkable progress achieved, most neural architecture search (NAS) methods focus on searching for one single accurate and robust architecture. To further build models with better generalization capability and performance, model ensemble is usually adopted and performs better than stand-alone models. Inspired by the merits of model ensemble, we propose to search for multiple diverse models simultaneously as an alternative way to find powerful models. Searching for ensembles is non-trivial and has two key challenges: enlarged search space and potentially more complexity for the searched model. In this paper, we propose a one-shot neural ensemble architecture search (NEAS) solution that addresses the two challenges. For the first challenge, we introduce a novel diversity-based metric to guide search space shrinking, considering both the potentiality and diversity of candidate operators. For the second challenge, we enable a new search dimen

sion to learn layer sharing among different models for efficiency purposes. The
experiments on ImageNet clearly demonstrate that our solution can improve the su
pernet's capacity of ranking ensemble architectures, and further lead to better
search results. The discovered architectures achieve superior performance compar
ed with state-of-the-arts such as MobileNetV3 and EfficientNet families under al
igned settings. Moreover, we evaluate the generalization ability and robustness
of our searched architecture on the COCO detection benchmark and achieve a 3.1%
improvement on AP compared with MobileNetV3. Codes and models are available here
.
****
Disentangled Cycle Consistency for Highly-Realistic Virtual Try-On
Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, Ping Luo; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 16928-16937
Image virtual try-on replaces the clothes on a person image with a desired in-sh
op clothes image. It is challenging because the person and the in-shop clothes a
re unpaired. Existing methods formulate virtual try-on as either in-painting or
cycle consistency. Both of these two formulations encourage the generation netwo
rks to reconstruct the input image in a self-supervised manner. However, existin
g methods do not differentiate clothing and non-clothing regions. A straightforw
ard generation impedes the virtual try-on quality because of the heavily coupled
 image contents. In this paper, we propose a Disentangled Cycle-consistency Try-
On Network (DCTON). The DCTON is able to produce highly-realistic try-on images
by disentangling important components of virtual try-on including clothes warpin
g, skin synthesis, and image composition. Moreover, DCTON can be naturally train
ed in a self-supervised manner following cycle consistency learning. Extensive e
xperiments on challenging benchmarks show that DCTON outperforms state-of-the-ar
t approaches favorably.
****
M3DSSD: Monocular 3D Single Stage Object Detector
Shujie Luo, Hang Dai, Ling Shao, Yong Ding; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6145-6154
In this paper, we propose a Monocular 3D Single Stage object Detector (M3DSSD) w
ith feature alignment and asymmetric non-local attention. Current anchor-based m
onocular 3D object detection methods suffer from feature mismatching. To overcom
e this, we propose a two-step feature alignment approach. In the first step, the
 shape alignment is performed to enable the receptive field of the feature map t
o focus on the pre-defined anchors with high confidence scores. In the second st
ep, the center alignment is used to align the features at 2D/3D centers. Further
, it is often difficult to learn global information and capture long-range relat
ionships, which are important for the depth prediction of objects. Therefore, we
 propose a novel asymmetric non-local attention block with multi-scale sampling
to extract depth-wise features. The proposed M3DSSD achieves significantly bette
r performance than the monocular 3D object detection methods on the KITTI datase
t, in both 3D object detection and bird's eye view tasks. The code is released a
t https://github.com/mumianyuxin/M3DSSD.
****
Structure-Aware Face Clustering on a Large-Scale Graph With 107 Nodes
Shuai Shen, Wanhua Li, Zheng Zhu, Guan Huang, Dalong Du, Jiwen Lu, Jie Zhou; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2021, pp. 9085-9094
Face clustering is a promising method for annotating unlabeled face images. Rece
nt supervised approaches have boosted the face clustering accuracy greatly, howe
ver their performance is still far from satisfactory. These methods can be rough
ly divided into global-based and local-based ones. Global-based methods suffer f
rom the limitation of training data scale, while local-based ones are difficult
to grasp the whole graph structure information and usually take a long time for
inference. Previous approaches fail to tackle these two challenges simultaneousl
y. To address the dilemma of large-scale training and efficient inference, we pr
opose the STructure-AwaRe Face Clustering (STAR-FC) method. Specifically, we des

ign a structure-preserved subgraph sampling strategy to explore the power of lar
ge-scale training data, which can increase the training data scale from 10^5 to
10^7. During inference, the STAR-FC performs efficient full-graph clustering wit
h two steps: graph parsing and graph refinement. And the concept of node intimac
y is introduced in the second step to mine the local structural information. The
 STAR-FC gets 91.97 pairwise F-score on partial MS1M within 310s which surpasses
 the state-of-the-arts. Furthermore, we are the first to train on very large-sca
le graph with 20M nodes, and achieve superior inference results on 12M testing d
ata. Overall, as a simple and effective method, the proposed STAR-FC provides a
strong baseline for large-scale face clustering. Code is available at https://ss
tzal.github.io/STAR-FC/.
****
Objects Are Different: Flexible Monocular 3D Object Detection
Yunpeng Zhang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2021, pp. 3289-3298
The precise localization of 3D objects from a single image without depth informa
tion is a highly challenging problem. Most existing methods adopt the same appro
ach for all objects regardless of their diverse distributions, leading to limite
d performance especially for truncated objects. In this paper, we propose a flex
ible framework for monocular 3D object detection which explicitly decouples the
truncated objects and adaptively combines multiple approaches for object depth e
stimation. Specifically, we decouple the edge of the feature map for predicting
long-tail truncated objects so that the optimization of normal objects is not in
fluenced. Furthermore, we formulate the object depth estimation as an uncertaint
y-guided ensemble of directly regressed object depth and solved depths from diff
erent groups of keypoints. Experiments demonstrate that our method outperforms t
he state-of-the-art method by relatively 27% for moderate level and 30% for hard
 level in the test set of KITTI benchmark while maintaining real-time efficiency
.
****
Permuted AdaIN: Reducing the Bias Towards Global Statistics in Image Classificat
ion
Oren Nuriel, Sagie Benaim, Lior Wolf; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9482-9491
Recent work has shown that convolutional neural network classifiers overly rely
on texture at the expense of shape cues. We make a similar but different distinc
tion between shape and local image cues, on the one hand, and global image stati
stics, on the other. Our method, called Permuted Adaptive Instance Normalization
 (pAdaIN), reduces the representation of global statistics in the hidden layers
of image classifiers. pAdaIN samples a random permutation p that rearranges the
samples in a given batch. Adaptive Instance Normalization (AdaIN) is then applie
d between the activations of each (non-permuted) sample i and the corresponding
activations of the sample p(i), thus swapping statistics between the samples of
the batch. Since the global image statistics are distorted, this swapping proced
ure causes the network to rely on cues, such as shape or texture. By choosing th
e random permutation with probability p and the identity permutation otherwise,
one can control the effect's strength. With the correct choice of p, fixed aprio
ri for all experiments and selected without considering test data, our method co
nsistently outperforms baselines in multiple settings. In image classification,
our method improves on both CIFAR100 and ImageNet using multiple architectures.
In the setting of robustness, our method improves on both ImageNet-C and Cifar-1
00-C for multiple architectures. In the setting of domain adaptation and domain
generalization, our method achieves state of the art results on the transfer lea
rning task from GTAV to Cityscapes and on the PACS benchmark.
****
Pixel Codec Avatars
Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De la T
orre, Yaser Sheikh; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2021, pp. 64-73
Telecommunication with photorealistic avatars in virtual or augmented reality is

 a promising path for achieving authentic face-to-face communication in 3D over remote physical distances. In this work, we present the Pixel Codec Avatars (PiCA): a deep generative model of 3D human faces that achieves state of the art reconstruction performance while being computationally efficient and adaptive to the rendering conditions during execution. Our model combines two core ideas: (1) a fully convolutional architecture for decoding spatially varying features, and (2) a rendering-adaptive per-pixel decoder. Both techniques are integrated via a dense surface representation that is learned in a weakly-supervised manner from low-topology mesh tracking over training images. We demonstrate that PiCA improves reconstruction over existing techniques across testing expressions and views on persons of different gender and skin tone. Importantly, we show that the PiCA model is much smaller than the state-of-art baseline model, and makes multi-person telecommunicaiton possible: on a single Oculus Quest 2 mobile VR headset, 5 avatars are rendered in realtime in the same scene.
****
SimPLE: Similar Pseudo Label Exploitation for Semi-Supervised Classification
Zijian Hu, Zhengyu Yang, Xuefeng Hu, Ram Nevatia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15099-15108
A common classification task situation is where one has a large amount of data available for training, but only a small portion is annotated with class labels. The goal of semi-supervised training, in this context, is to improve classification accuracy by leverage information not only from labeled data but also from a large amount of unlabeled data. Recent works have developed significant improvements by exploring the consistency constrain between differently augmented labeled and unlabeled data. Following this path, we propose a novel unsupervised objective that focuses on the less studied relationship between the high confidence unlabeled data that are similar to each other. The new proposed Pair Loss minimizes the statistical distance between high confidence pseudo labels with similarity above a certain threshold. Combining the Pair Loss with the techniques developed by the MixMatch family, our proposed SimPLE algorithm shows significant performance gains over previous algorithms on CIFAR-100 and Mini-ImageNet, and is on par with the state-of-the-art methods on CIFAR-10 and SVHN. Furthermore, SimPLE also outperforms the state-of-the-art methods in the transfer learning setting, where models are initialized by the weights pre-trained on ImageNet or DomainNet-Real. The code is available at github.com/zijian-hu/SimPLE.
****
Context-Aware Layout to Image Generation With Enhanced Object Appearance
Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, Tao Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15049-15058
A layout to image (L2I) generation model aims to generate a complicated image containing multiple objects (things) against natural background (stuff), conditioned on a given layout. Built upon the recent advances in generative adversarial networks (GANs), recent L2I models have made great progress. However, a close inspection of their generated images reveals two major limitations: (1) the object-to-object as well as object-to-stuff relations are often broken and (2) each object's appearance is typically distorted lacking the key defining characteristics associated with the object class. We argue that these are caused by the lack of context-aware object and stuff feature encoding in their generators, and location-sensitive appearance representation in their discriminators. To address these limitations, two new modules are proposed in this work. First, a contextual feature transformation module is introduced in the generator to ensure that the generated feature encoding of either object or stuff is aware of other co-existing objects/stuff in the scene. Second, instead of feeding location-insensitive image features to the discriminator, we use the Gram matrix computed from the feature maps of the generated object images to preserve location-sensitive information, resulting in much enhanced object appearance. Extensive experiments show that the proposed method achieves state-of-the-art performance on the COCO-Thing-Stuff and Visual Genome benchmarks.

****
Mask-Embedded Discriminator With Region-Based Semantic Regularization for Semi-Supervised Class-Conditional Image Synthesis
Yi Liu, Xiaoyang Huo, Tianyi Chen, Xiangping Zeng, Si Wu, Zhiwen Yu, Hau-San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5506-5515
Semi-supervised generative learning (SSGL) makes use of unlabeled data to achieve a trade-off between the data collection/annotation effort and generation performance, when adequate labeled data are not available. Learning precise class semantics is crucial for class-conditional image synthesis with limited supervision. Toward this end, we propose a semi-supervised Generative Adversarial Network with a Mask-Embedded Discriminator, which is referred to as MED-GAN. By incorporating a mask embedding module, the discriminator features are associated with spatial information, such that the focus of the discriminator can be limited in the specified regions when distinguishing between real and synthesized images. A generator is enforced to synthesize the instances holding more precise class semantics in order to deceive the enhanced discriminator. Also benefiting from mask embedding, region-based semantic regularization is imposed on the discriminator feature space, and the degree of separation between real and fake classes and among object categories can thus be increased. This eventually improves class-conditional distribution matching between real and synthesized data. In the experiments, the superior performance of MED-GAN demonstrates the effectiveness of mask embedding and associated regularizers in facilitating SSGL.
****
LEAP: Learning Articulated Occupancy of People
Marko Mihajlovic, Yan Zhang, Michael J. Black, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10461-10471
Substantial progress has been made on modeling rigid 3D objects using deep implicit representations. Yet, extending these methods to learn neural models of human shape is still in its infancy. Human bodies are complex and the key challenge is to learn a representation that generalizes such that it can express body shape deformations for unseen subjects in unseen, highly-articulated, poses. To address this challenge, we introduce LEAP (LEarning Articulated occupancy of People), a novel neural occupancy representation of the human body. Given a set of bone transformations (i.e. joint locations and rotations) and a query point in space, LEAP first maps the query point to a canonical space via learned linear blend skinning (LBS) functions and then efficiently queries the occupancy value via an occupancy network that models accurate identity- and pose-dependent deformations in the canonical space. Experiments show that our canonicalized occupancy estimation with the learned LBS functions greatly improves the generalization capability of the learned occupancy representation across various human shapes and poses, outperforming existing solutions in all settings.
****
ANR: Articulated Neural Rendering for Virtual Avatars
Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, Christoph Lassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3722-3731
Deferred Neural Rendering (DNR) uses a three-step pipeline to translate a mesh representation into an RGB image. The combination of a traditional rendering stack with neural networks hits a sweet spot in terms of computational complexity and realism of the resulting images. Using skinned meshes for animatable objects is a natural extension for the framework and would open it up to a plethora of applications. However, in this case the neural shading step must account for deformations that are possibly not captured in the mesh, as well as alignment accuracies and dynamics---which is not well-supported in the DNR pipeline. In this paper, we present an in-depth study of possibilities to develop the DNR framework towards handling these cases. We outline several steps that can be easily integrated into the DNR pipeline for addressing stability and deformation. We demonstrate their efficiency by building a virtual avatar pipeline, a highly challenging c

ase with animation and clothing deformation, and show the superiority of the presented method not only with respect to the DNR pipeline but also with methods specifically for virtual avatar creation and animation. In two user studies, we observe a clear preference for our avatar model and outperform other methods on SSIM and LPIPS metrics. Perceptually, we observe better temporal stability, level of detail and plausibility.
****
Flow-Based Kernel Prior With Application to Blind Super-Resolution
Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10601-10610
Kernel estimation is generally one of the key problems for blind image super-resolution (SR). Recently, Double-DIP proposes to model the kernel via a network architecture prior, while KernelGAN employs the deep linear network and several regularization losses to constrain the kernel space. However, they fail to fully exploit the general SR kernel assumption that anisotropic Gaussian kernels are sufficient for image SR. To address this issue, this paper proposes a normalizing flow-based kernel prior (FKP) for kernel modeling. By learning an invertible mapping between the anisotropic Gaussian kernel distribution and a tractable latent distribution, FKP can be easily used to replace the kernel modeling modules of Double-DIP and KernelGAN. Specifically, FKP optimizes the kernel in the latent space rather than the network parameter space, which allows it to generate reasonable kernel initialization, traverse the learned kernel manifold and improve the optimization stability. Extensive experiments on synthetic and real-world images demonstrate that the proposed FKP can significantly improve the kernel estimation accuracy with less parameters, runtime and memory usage, leading to state-of-the-art blind SR results.
****
Probabilistic Selective Encryption of Convolutional Neural Networks for Hierarchical Services
Jinyu Tian, Jiantao Zhou, Jia Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2205-2214
Model protection is vital when deploying Convolutional Neural Networks (CNNs) for commercial services, due to the massive costs of training them. In this work, we propose a selective encryption (SE) algorithm to protect CNN models from unauthorized access, with a unique feature of providing hierarchical services to users. Our algorithm firstly selects important model parameters via the proposed Probabilistic Selection Strategy (PSS). It then encrypts the most important parameters with the designed encryption method called Distribution Preserving Random Mask (DPRM), so as to maximize the performance degradation by encrypting only a very small portion of model parameters. We also design a set of access permissions, using which different amount of most important model parameters can be decrypted. Hence, different levels of model performance can be naturally provided for users. Experimental results demonstrate that the proposed scheme could effectively protect the classification model VGG19 by merely encrypting 8% parameters of convolutional layers. We also implement the proposed model protection scheme in the denoising model DnCNN, showcasing the hierarchical denoising services.
****
Cuboids Revisited: Learning Robust 3D Shape Fitting to Single RGB Images
Florian Kluger, Hanno Ackermann, Eric Brachmann, Michael Ying Yang, Bodo Rosenhahn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13070-13079
Humans perceive and construct the surrounding world as an arrangement of simple parametric models. In particular, man-made environments commonly consist of volumetric primitives such as cuboids or cylinders. Inferring these primitives is an important step to attain high-level, abstract scene descriptions. Previous approaches directly estimate shape parameters from a 2D or 3D input, and are only able to reproduce simple objects, yet unable to accurately parse more complex 3D scenes. In contrast, we propose a robust estimator for primitive fitting, which can meaningfully abstract real-world environments using cuboids. A RANSAC estimat

or guided by a neural network fits these primitives to 3D features, such as a depth map. We condition the network on previously detected parts of the scene, thus parsing it one-by-one. To obtain 3D features from a single RGB image, we additionally optimise a feature extraction CNN in an end-to-end manner. However, naively minimising point-to-primitive distances leads to large or spurious cuboids occluding parts of the scene behind. We thus propose an occlusion-aware distance metric correctly handling opaque scenes. The proposed algorithm does not require labour-intensive labels, such as cuboid annotations, for training. Results on the challenging NYU Depth v2 dataset demonstrate that the proposed algorithm successfully abstracts cluttered real-world 3D scene layouts.
****
Dive Into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition
Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6248-6257
Due to the subjective annotation and the inherent inter-class similarity of facial expressions, one of key challenges in Facial Expression Recognition (FER) is the annotation ambiguity. In this paper, we proposes a solution, named DMUE, to address the problem of annotation ambiguity from two perspectives: the latent Distribution Mining and the pairwise Uncertainty Estimation. For the former, an auxiliary multi-branch learning framework is introduced to better mine and describe the latent distribution in the label space. For the latter, the pairwise relationship of semantic feature between instances are fully exploited to estimate the ambiguity extent in the instance space. The proposed method is independent to the backbone architectures, and brings no extra burden for inference. The experiments are conducted on the popular real-world benchmarks and the synthetic noisy datasets. Either way, the proposed DMUE stably achieves leading performance.
****
Attention-Guided Image Compression by Deep Reconstruction of Compressive Sensed Saliency Skeleton
Xi Zhang, Xiaolin Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13354-13364
We propose a deep learning system for attention-guided dual-layer image compression (AGDL). In the AGDL compression system, an image is encoded into two layers, a base layer and an attention-guided refinement layer. Unlike the existing ROI image compression methods that spend an extra bit budget equally on all pixels in ROI, AGDL employs a CNN module to predict those pixels on and near a saliency sketch within ROI that are critical to perceptual quality. Only the critical pixels are further sampled by compressive sensing (CS) to form a very compact refinement layer. Another novel CNN method is developed to jointly decode the two compression code layers for a much refined reconstruction, while strictly satisfying the transmitted CS constraints on perceptually critical pixels. Extensive experiments demonstrate that the proposed AGDL system advances the state of the art in perception-aware image compression.
****
Cluster-Wise Hierarchical Generative Model for Deep Amortized Clustering
Huafeng Liu, Jiaqi Wang, Liping Jing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15109-15118
In this paper, we propose Cluster-wise Hierarchical Generative Model for deep amortized clustering (CHiGac). It provides an efficient neural clustering architecture by grouping data points in a cluster-wise view rather than point-wise view. CHiGac simultaneously learns what makes a cluster, how to group data points into clusters, and how to adaptively control the number of clusters. The dedicated cluster generative process is able to sufficiently exploit pair-wise or higher-order interactions between data points in both inter- and intra-cluster, which is useful to sufficiently mine the hidden structure among data. To efficiently minimize the generalized lower bound of CHiGac, we design an Ergodic Amortized Inference (EAI) strategy by considering the average behavior over sequence on an inner variational parameter trajectory, which is theoretically proven to reduce the

amortization gap. A series of experiments have been conducted on both synthetic and real-world data. The experimental results demonstrated that CHiGac can efficiently and accurately cluster datasets in terms of both internal and external evaluation metrics (DBI and ACC).
****
Mirror3D: Depth Refinement for Mirror Surfaces
Jiaqi Tan, Weijie Lin, Angel X. Chang, Manolis Savva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15990-15999
Despite recent progress in depth sensing and 3D reconstruction, mirror surfaces are a significant source of errors. To address this problem, we create the Mirror3D dataset: a 3D mirror plane dataset based on three RGBD datasets (Matterpot3D, NYUv2 and ScanNet) containing 7,011 mirror instance masks and 3D planes. We then develop Mirror3DNet: a module that refines raw sensor depth or estimated depth to correct errors on mirror surfaces. Our key idea is to estimate the 3D mirror plane based on RGB input and surrounding depth context, and use this estimate to directly regress mirror surface depth. Our experiments show that Mirror3DNet significantly mitigates errors from a variety of input depth data, including raw sensor depth and depth estimation or completion methods.
****
Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning
Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, Han Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16684-16693
Contrastive learning methods for unsupervised visual representation learning have reached remarkable levels of transfer performance. We argue that the power of contrastive learning has yet to be fully unleashed, as current methods are trained only on instance-level pretext tasks, leading to representations that may be sub-optimal for downstream tasks requiring dense pixel predictions. In this paper, we introduce pixel-level pretext tasks for learning dense feature representations. The first task directly applies contrastive learning at the pixel level. We additionally propose a pixel-to-propagation consistency task that produces better results, even surpassing the state-of-the-art approaches by a large margin. Specifically, it achieves 60.2 AP, 41.4 / 40.5 mAP and 77.2 mIoU when transferred to Pascal VOC object detection (C4), COCO object detection (FPN / C4) and Cityscapes semantic segmentation using a ResNet-50 backbone network, which are 2.6 AP, 0.8 / 1.0 mAP and 1.0 mIoU better than the previous best methods built on instance-level contrastive learning. Moreover, the pixel-level pretext tasks are found to be effective for pre-training not only regular backbone networks but also head networks used for dense downstream tasks, and are complementary to instance-level contrastive methods. These results demonstrate the strong potential of defining pretext tasks at the pixel level, and suggest a new path forward in unsupervised visual representation learning. Code is available at https://github.com/zdaxie/PixPro.
****
Reciprocal Transformations for Unsupervised Video Object Segmentation
Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15455-15464
Unsupervised video object segmentation (UVOS) aims at segmenting the primary objects in videos without any human intervention. Due to the lack of prior knowledge about the primary objects, identifying them from videos is the major challenge of UVOS. Previous methods often regard the moving objects as primary ones and rely on optical flow to capture the motion cues in videos, but the flow information alone is insufficient to distinguish the primary objects from the background objects that move together. This is because, when the noisy motion features are combined with the appearance features, the localization of the primary objects is misguided. To address this problem, we propose a novel reciprocal transformation network to discover primary objects by correlating three key factors: the int

ra-frame contrast, the motion cues, and temporal coherence of recurring objects. Each corresponds to a representative type of primary object, and our reciprocal mechanism enables an organic coordination of them to effectively remove ambiguous distractions from videos. Additionally, to exclude the information of the moving background objects from motion features, our transformation module enables to reciprocally transform the appearance features to enhance the motion features, so as to focus on the moving objects with salient appearance while removing the co-moving outliers. Experiments on the public benchmarks demonstrate that our model significantly outperforms the state-of-the-art methods.
****
Detection, Tracking, and Counting Meets Drones in Crowds: A Benchmark
Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, Siwei Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7812-7821
To promote the developments of object detection, tracking and counting algorithms in drone-captured videos, we construct a benchmark with a new drone-captured large-scale dataset, named as DroneCrowd, formed by 112 video clips with 33,600 HD frames in various scenarios. Notably, we annotate 20,800 people trajectories with 4.8 million heads and several video-level attributes. Meanwhile, we design the Space-Time Neighbor-Aware Network (STNNet) as a strong baseline to solve object detection, tracking and counting jointly in dense crowds. STNNet is formed by the feature extraction module, followed by the density map estimation heads, and localization and association subnets. To exploit the context information of neighboring objects, we design the neighboring context loss to guide the association subnet training, which enforces consistent relative position of nearby objects in temporal domain. Extensive experiments on our DroneCrowd dataset demonstrate that STNNet performs favorably against the state-of-the-arts.
****
Learning Complete 3D Morphable Face Models From Images and Videos
Mallikarjun B R, Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3361-3371
Most 3D face reconstruction methods rely on 3D morphable models, which disentangle the space of facial deformations into identity and expression geometry, and skin reflectance. These models are typically learned from a limited number of 3D scans and thus do not generalize well across different identities and expressions. We present the first approach to learn complete 3D models of face identity and expression geometry, and reflectance, just from images and videos. The virtually endless collection of such data, in combination with our self-supervised learning-based approach allows for learning face models that generalize beyond the span of existing approaches. Our network design and loss functions ensure a disentangled parameterization of not only identity and albedo, but also, for the first time, an expression basis. Our method also allows for in-the-wild monocular reconstruction at test time. We show that our learned models better generalize and lead to higher quality image-based reconstructions than existing approaches. We show that the learned model can also be personalized to a video, for a better capture of the geometry and albedo.
****
Bottom-Up Shift and Reasoning for Referring Image Segmentation
Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11266-11275
Referring image segmentation aims to segment the referent that is the corresponding object or stuff referred by a natural language expression in an image. Its main challenge lies in how to effectively and efficiently differentiate between the referent and other objects of the same category as the referent. In this paper, we tackle the challenge by jointly performing compositional visual reasoning and accurate segmentation in a single stage via the proposed novel Bottom-Up Shift (BUS) and Bidirectional Attentive Refinement (BIAR) modules. Specifically, BUS progressively locates the referent along hierarchical reasoning steps implied

by the expression. At each step, it locates the corresponding visual region by disambiguating between similar regions, where the disambiguation bases on the relationships between regions. By the explainable visual reasoning, BUS explicitly aligns linguistic components with visual regions so that it can identify all the mentioned entities in the expression. BIAR fuses multi-level features via a two-way attentive message passing, which captures the visual details relevant to the referent to refine segmentation results. Experimental results demonstrate that the proposed method consisting of BUS and BIAR modules, can not only consistently surpass all existing state-of-the-art algorithms across common benchmark datasets but also visualize interpretable reasoning steps for stepwise segmentation. Code is available at https://github.com/incredibleXM/BUSNet.
****
Sparse Auxiliary Networks for Unified Monocular Depth Prediction and Completion
Vitor Guizilini, Rares Ambrus, Wolfram Burgard, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11078-11088
Estimating scene geometry from cost-effective sensors is key for robots. In this paper, we study the problem of predicting dense depth from a single RGB image (monodepth) with optional sparse measurements from low-cost active depth sensors. We introduce Sparse Auxiliary Networks (SAN), a new module enabling monodepth networks to perform both the tasks of depth prediction and completion, depending on whether only RGB images or also sparse point clouds are available at inference time. First, we decouple the image and depth map encoding stages using sparse convolutions to process only the valid depth map pixels. Second, we inject this information, when available, into the skip connections of the depth prediction network, augmenting its features. Through extensive experimental analysis on one indoor (NYUv2) and two outdoor (KITTI and DDAD) benchmarks, we demonstrate that our proposed SAN architecture is able to simultaneously learn both tasks, while achieving a new state of the art in depth prediction by a significant margin.
****
DeepMetaHandles: Learning Deformation Meta-Handles of 3D Meshes With Biharmonic Coordinates
Minghua Liu, Minhyuk Sung, Radomir Mech, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12-21
We propose DeepMetaHandles, a 3D conditional generative model based on mesh deformation. Given a collection of 3D meshes of a category and their deformation handles (control points), our method learns a set of meta-handles for each shape, which are represented as combinations of the given handles. The disentangled meta-handles factorize all the plausible deformations of the shape, while each of them corresponds to an intuitive deformation. A new deformation can then be generated by sampling the coefficients of the meta-handles in a specific range. We employ biharmonic coordinates as the deformation function, which can smoothly propagate the control points' translations to the entire mesh. To avoid learning zero deformation as meta-handles, we incorporate a target-fitting module which deforms the input mesh to match a random target. To enhance deformations' plausibility, we employ a soft-rasterizer-based discriminator that projects the meshes to a 2D space. Our experiments demonstrate the superiority of the generated deformations as well as the interpretability and consistency of the learned meta-handles.
****
Panoptic Segmentation Forecasting
Colin Graber, Grace Tsai, Michael Firman, Gabriel Brostow, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12517-12526
Our goal is to forecast the near future given a set of recent observations. We think this ability to forecast, i.e., to anticipate, is integral for the success of autonomous agents which need not only passively analyze an observation but also must react to it in real-time. Importantly, accurate forecasting hinges upon the chosen scene decomposition. We think that superior forecasting can be achieved by decomposing a dynamic scene into individual 'things' and background 'stuff

'. Background 'stuff' largely moves because of camera motion, while foreground ' things' move because of both camera and individual object motion. Following this decomposition, we introduce panoptic segmentation forecasting. Panoptic segmentation forecasting opens up a middle-ground between existing extremes, which either forecast instance trajectories or predict the appearance of future image frames. To address this task we develop a two-component model: one component learns the dynamics of the background stuff by anticipating odometry, the other one anticipates the dynamics of detected things. We establish a leaderboard for this novel task, and validate a state-of-the-art model that outperforms available baselines.
****
SRDAN: Scale-Aware and Range-Aware Domain Adaptation Network for Cross-Dataset 3D Object Detection
Weichen Zhang, Wen Li, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6769-6779
Geometric characteristic plays an important role in the representation of an object in 3D point clouds. For example, large objects often contain more points, while small ones contain fewer points. The point clouds of objects near the capture device are denser, while those of distant objects are sparser. These issues bring new challenges to 3D object detection, especially under the domain adaptation scenarios. In this work, we propose a new cross-dataset 3D object detection method named Scale-aware and Range-aware Domain Adaptation Network (SRDAN). We take advantage of the geometric characteristics of 3D data (i.e., size and distance), and propose the scale-aware domain alignment and the range-aware domain alignment strategies to guide the distribution alignment between two domains. For scale-aware domain alignment, we design a 3D voxel-based feature pyramid network to extract multi-scale semantic voxel features, and align the features and instances with similar scales between two domains. For range-aware domain alignment, we introduce a range-guided domain alignment module to align the features of objects according to their distance to the capture device. Extensive experiments under three different scenarios demonstrate the effectiveness of our SRDAN approach, and comprehensive ablation study also validates the importance of geometric characteristics for cross-dataset 3D object detection.
****
Pedestrian and Ego-Vehicle Trajectory Prediction From Monocular Camera
Lukas Neumann, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10204-10212
Predicting future pedestrian trajectory is a crucial component of autonomous driving systems, as recognizing critical situations based only on current pedestrian position may come too late for any meaningful corrective action (e.g. breaking) to take place. In this paper, we propose a new method to predict future position of pedestrians, with respect to a predicted future position of the ego-vehicle, thus giving a assistive/autonomous driving system sufficient time to respond. The method explicitly disentangles actual movement of pedestrians in real world from the ego-motion of the vehicle, using a future pose prediction network trained in self-supervised fashion, which allows the method to observe and predict the intrinsic pedestrian motion in a normalised view, that captures the same real-world location across multiple frames. The method is evaluated on two public datasets, where it achieves state-of-the-art results in pedestrian trajectory prediction from an on-board camera.
****
Globally Optimal Relative Pose Estimation With Gravity Prior
Yaqing Ding, Daniel Barath, Jian Yang, Hui Kong, Zuzana Kukelova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 394-403
Smartphones, tablets and camera systems used, e.g., in cars and UAVs, are typically equipped with IMUs (inertial measurement units) that can measure the gravity vector accurately. Using this additional information, the y-axes of the cameras can be aligned, reducing their relative orientation to a single degree-of-freedom. With this assumption, we propose a novel globally optimal solver, minimizing

 the algebraic error in the least squares sense, to estimate the relative pose i
n the over-determined case. Based on the epipolar constraint, we convert the opt
imization problem into solving two polynomials with only two unknowns. Also, a f
ast solver is proposed using the first-order approximation of the rotation. The
proposed solvers are compared with the state-of-the-art ones on four real-world
datasets with approx. 50000 image pairs in total. Moreover, we collected a datas
et, by a smartphone, consisting of 10933 image pairs, gravity directions and gro
und truth 3D reconstructions. The source code and dataset are available at https
://github.com/yaqding/opt_pose_gravity
****
Mutual CRF-GNN for Few-Shot Learning
Shixiang Tang, Dapeng Chen, Lei Bai, Kaijian Liu, Yixiao Ge, Wanli Ouyang; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2021, pp. 2329-2339
Graph-neural-networks (GNN) is a rising trend for few-shot learning. A critical
component in GNN is the affinity. Typically, affinity in GNN is mainly computed
in the feature space, e.g., pairwise features, and does not take fully advantage
 of semantic labels associated to these features. In this paper, we propose a no
vel Mutual CRF-GNN (MCGN). In this MCGN, the labels and features of support data
 are used by the CRF for inferring GNN affinities in a principled and probabilis
tic way. Specifically, we construct a Conditional Random Field (CRF) conditioned
 on labels and features of support data to infer a affinity in the label space.
Such affinity is fed to the GNN as the node-wise affinity. GNN and CRF mutually
contributes to each other in MCGN. For GNN, CRF provides valuable affinity infor
mation. For CRF, GNN provides better features for inferring affinity. Experiment
al results show that our approach outperforms state-of-the-arts on datasets mini
ImageNet, tieredImageNet, and CIFAR-FS on both 5-way 1-shot and 5-way 5-shot set
tings.
****
Weakly Supervised Action Selection Learning in Video
Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, Guangwei Yu; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp
. 7587-7596
Localizing actions in video is a core task in computer vision. The weakly superv
ised temporal localization problem investigates whether this task can be adequat
ely solved with only video-level labels, significantly reducing the amount of ex
pensive and error-prone annotation that is required. A common approach is to tra
in a frame-level classifier where frames with the highest class probability are
selected to make a video-level prediction. Frame-level activations are then used
 for localization. However, the absence of frame-level annotations cause the cla
ssifier to impart class bias on every frame. To address this, we propose the Act
ion Selection Learning (ASL) approach to capture the general concept of action,
a property we refer to as "actionness". Under ASL, the model is trained with a n
ovel class-agnostic task to predict which frames will be selected by the classif
ier. Empirically, we show that ASL outperforms leading baselines on two popular
benchmarks THUMOS-14 and ActivityNet-1.2, with 10.3% and 5.7% relative improveme
nt respectively. We further analyze the properties of ASL and demonstrate the im
portance of actionness. Full code for this work is available here https://github
.com/layer6ai-labs/ASL
****
Learning Student Networks in the Wild
Hanting Chen, Tianyu Guo, Chang Xu, Wenshuo Li, Chunjing Xu, Chao Xu, Yunhe Wang
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2021, pp. 6428-6437
Data-free learning for student networks is a new paradigm for solving users' anx
iety caused by the privacy problem of using original training data. Since the ar
chitectures of modern convolutional neural networks (CNNs) are compact and sophi
sticated, the alternative images or meta-data generated from the teacher network
 are often broken. Thus, the student network cannot achieve the comparable perfo
rmance to that of the pre-trained teacher network especially on the large-scale

image dataset. Different to previous works, we present to maximally utilize the massive available unlabeled data in the wild. Specifically, we first thoroughly analyze the output differences between teacher and student network on the original data and develop a data collection method. Then, a noisy knowledge distillation algorithm is proposed for achieving the performance of the student network. In practice, an adaptation matrix is learned with the student network for correcting the label noise produced by the teacher network on the collected unlabeled images. The effectiveness of our DFND (Data-Free Noisy Distillation) method is then verified on several benchmarks to demonstrate its superiority over state-of-the-art data-free distillation methods. Experiments on various datasets demonstrate that the student networks learned by the proposed method can achieve comparable performance with those using the original dataset. Code is available at https://github.com/huawei-noah/Data-Efficient-Model-Compression
****
Distilling Knowledge via Knowledge Review
Pengguang Chen, Shu Liu, Hengshuang Zhao, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5008-5017
Knowledge distillation transfers knowledge from the teacher network to the student one, with the goal of greatly improving the performance of the student network. Previous methods mostly focus on proposing feature transformation and loss functions between the same level's features to improve the effectiveness. We differently study the factor of connection path cross levels between teacher and student networks, and reveal its great importance. For the first time in knowledge distillation, cross-stage connection paths are proposed. A new review mechanism becomes vastly effective and structurally simple. Our finally designed nested and compact framework requires negligible computation overhead, and outperforms other methods on a variety of tasks. We apply our method to classification, object detection, and instance segmentation tasks. All of them witness significant student network performance improvement.
****
DoDNet: Learning To Segment Multi-Organ and Tumors From Multiple Partially Labeled Datasets
Jianpeng Zhang, Yutong Xie, Yong Xia, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1195-1204
Due to the intensive cost of labor and expertise in annotating 3D medical images at a voxel level, most benchmark datasets are equipped with the annotations of only one type of organs and/or tumors, resulting in the so-called partially labeling issue. To address this issue, we propose a dynamic on-demand network (DoDNet) that learns to segment multiple organs and tumors on partially labeled datasets. DoDNet consists of a shared encoder-decoder architecture, a task encoding module, a controller for dynamic filter generation, and a single but dynamic segmentation head. The information of current segmentation task is encoded as a task-aware prior to tell the model what the task is expected to achieve. Different from existing approaches which fix kernels after training, the kernels in dynamic head are generated adaptively by the controller, conditioned on both input image and assigned task. Thus, DoDNet is able to segment multiple organs and tumors, as done by multiple networks or a multi-head network, in a much efficient and flexible manner. We created a large-scale partially labeled dataset called MOTS and demonstrated the superior performance of our DoDNet over other competitors on seven organ and tumor segmentation tasks. We also transferred the weights pre-trained on MOTS to a downstream multi-organ segmentation task and achieved state-of-the-art performance. This study provides a general 3D medical image segmentation model that has been pre-trained on a large-scale partially labeled dataset and can be extended (after fine-tuning) to downstream volumetric medical data segmentation tasks. Code and models are available at https://git.io/DoDNet.
****
Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection
Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, Maja Pantic; Pr

oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5039-5049
Although current deep learning-based face forgery detectors achieve impressive performance in constrained scenarios, they are vulnerable to samples created by unseen manipulation methods. Some recent works show improvements in generalisation but rely on cues that are easily corrupted by common post-processing operations such as compression. In this paper, we propose LipForensics, a detection approach capable of both generalising to novel manipulations and withstanding various distortions. LipForensics targets high-level semantic irregularities in mouth movements, which are common in many generated videos. It consists in first pretraining a spatio-temporal network to perform visual speech recognition (lipreading), thus learning rich internal representations related to natural mouth motion. A temporal network is subsequently finetuned on fixed mouth embeddings of real and forged data in order to detect fake videos based on mouth movements without overfitting to low-level, manipulation-specific artefacts. Extensive experiments show that this simple approach significantly surpasses the state-of-the-art in terms of generalisation to unseen manipulations and robustness to perturbations, as well as shed light on the factors responsible for its performance.
****
Exploring Simple Siamese Representation Learning
Xinlei Chen, Kaiming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15750-15758
Siamese networks have become a common structure in various recent models for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to certain conditions for avoiding collapsing solutions. In this paper, we report surprising empirical results that simple Siamese networks can learn meaningful representations even using none of the following: (i) negative sample pairs, (ii) large batches, (iii) momentum encoders. Our experiments show that collapsing solutions do exist for the loss and structure, but a stop-gradient operation plays an essential role in preventing collapsing. We provide a hypothesis on the implication of stop-gradient, and further show proof-of-concept experiments verifying it. Our "SimSiam" method achieves competitive results on ImageNet and downstream tasks. We hope this simple baseline will motivate people to rethink the roles of Siamese architectures for unsupervised representation learning. Code is made available. (https://github.com/facebookresearch/simsiam)
****
CAMERAS: Enhanced Resolution and Sanity Preserving Class Activation Mapping for Image Saliency
Mohammad A. A. K. Jalwana, Naveed Akhtar, Mohammed Bennamoun, Ajmal Mian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16327-16336
Backpropagation image saliency aims at explaining model predictions by estimating model-centric importance of individual pixels in the input. However, class-insensitivity of the earlier layers in a network only allows saliency computation with low resolution activation maps of the deeper layers, resulting in compromised image saliency. Remedifying this can lead to sanity failures. We propose CAMERAS, a technique to compute high-fidelity backpropagation saliency maps without requiring any external priors and preserving the map sanity. Our method systematically performs multi-scale accumulation and fusion of the activation maps and backpropagated gradients to compute precise saliency maps. From accurate image saliency to articulation of relative importance of input features for different models, and precise discrimination between model perception of visually similar objects, our high-resolution mapping offers multiple novel insights into the black-box deep visual models, which are presented in the paper. We also demonstrate the utility of our saliency maps in adversarial setup by drastically reducing the norm of attack signals by focusing them on the precise regions identified by our maps. Our method also inspires new evaluation metrics and a sanity check for this developing research direction.
****

3D AffordanceNet: A Benchmark for Visual Object Affordance Understanding
Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1778-1787
The ability to understand the ways to interact with objects from visual cues, a.k.a. visual affordance, is essential to vision-guided robotic research. This involves categorizing, segmenting and reasoning of visual affordance. Relevant studies in 2D and 2.5D image domains have been made previously, however, a truly functional understanding of object affordance requires learning and prediction in the 3D physical domain, which is still absent in the community. In this work, we present a 3D AffordanceNet dataset, a benchmark of 23k shapes from 23 semantic object categories, annotated with 18 visual affordance categories. Based on this dataset, we provide three benchmarking tasks for evaluating visual affordance understanding, including full-shape, partial-view and rotation-invariant affordance estimations. Three state-of-the-art point cloud deep learning networks are evaluated on all tasks. In addition we also investigate a semi-supervised learning setup to explore the possibility to benefit from unlabeled data. Comprehensive results on our contributed dataset show the promise of visual affordance understanding as a valuable yet challenging benchmark.
****
Learning To Segment Actions From Visual and Language Instructions via Differentiable Weak Sequence Alignment
Yuhan Shen, Lu Wang, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10156-10165
We address the problem of unsupervised localization of key-steps and feature learning in instructional videos using both visual and language instructions. Our key observation is that the sequences of visual and linguistic key-steps are weakly aligned: there is an ordered one-to-one correspondence between most visual and language key-steps, while some key-steps in one modality are absent in the other. To recover the two sequences, we develop an ordered prototype learning module, which extracts visual and linguistic prototypes representing key-steps. On the other hand, to find weak alignment and perform feature learning, we develop a differentiable weak sequence alignment (DWSA) method that finds ordered one-to-one matching between sequences while allowing some items in a sequence to stay unmatched. We develop an efficient forward and backward algorithm for computing the alignment and the loss derivative with respect to parameters of visual and language feature learning modules. By experiments on two instructional video datasets, we show that our method significantly improves the state of the art.
****
Deep Implicit Templates for 3D Shape Representation
Zerong Zheng, Tao Yu, Qionghai Dai, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1429-1439
Deep implicit functions (DIFs), as a kind of 3D shape representation, are becoming more and more popular in the 3D vision community due to their compactness and strong representation power. However, unlike polygon mesh-based templates, it remains a challenge to reason dense correspondences or other semantic relationships across shapes represented by DIFs, which limits its applications in texture transfer, shape analysis and so on. To overcome this limitation and also make DIFs more interpretable, we propose Deep Implicit Templates, a new 3D shape representation that supports explicit correspondence reasoning in deep implicit representations. Our key idea is to formulate DIFs as conditional deformations of a template implicit function. To this end, we propose Spatial Warping LSTM, which decomposes the conditional spatial transformation into multiple point-wise transformations and guarantees generalization capability. Moreover, the training loss is carefully designed in order to achieve high reconstruction accuracy while learning a plausible template with accurate correspondences in an unsupervised manner. Experiments show that our method can not only learn a common implicit template for a collection of shapes, but also establish dense correspondences across all the shapes simultaneously without any supervision.
****

Semantic Image Matting
Yanan Sun, Chi-Keung Tang, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11120-11129
Natural image matting separates the foreground from background in fractional occupancy which can be caused by highly transparent objects, complex foreground (e.g., net or tree), and/or objects containing very fine details (e.g., hairs). Although conventional matting formulation can be applied to all of the above cases, no previous work has attempted to reason the underlying causes of matting due to various foreground semantics. We show how to obtain better alpha mattes by incorporating into our framework semantic classification of matting regions. Specifically, we consider and learn 20 classes of matting patterns, and propose to extend the conventional trimap to semantic trimap. The proposed semantic trimap can be obtained automatically through patch structure analysis within trimap regions. Meanwhile, we learn a multi-class discriminator to regularize the alpha prediction at semantic level, and content-sensitive weights to balance different regularization losses. Experiments on multiple benchmarks show that our method outperforms other methods and has achieved the most competitive state-of-the-art performance. Finally, we contribute a large-scale Semantic Image Matting Dataset with careful consideration of data balancing across different semantic classes. Code and dataset will be released.
****
Semi-Supervised Semantic Segmentation With Cross Pseudo Supervision
Xiaokang Chen, Yuhui Yuan, Gang Zeng, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2613-2622
In this paper, we study the semi-supervised semantic segmentation problem via exploring both labeled data and extra unlabeled data. We propose a novel consistency regularization approach, called cross pseudo supervision (CPS). Our approach imposes the consistency on two segmentation networks perturbed with different initialization for the same input image. The pseudo one-hot label map, output from one perturbed segmentation network, is used to supervise the other segmentation network with the standard cross-entropy loss, and vice versa. The CPS consistency has two roles: encourage high similarity between the predictions of two perturbed networks for the same input image, and expand training data by using the unlabeled data with pseudo labels.
****
Ranking Neural Checkpoints
Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, Boqing Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2663-2673
This paper is concerned with ranking many pre-trained deep neural networks (DNNs), called checkpoints, for the transfer learning to a downstream task. Thanks to the broad use of DNNs, we may easily collect hundreds of checkpoints from various sources. Which of them transfers the best to our downstream task of interest? Striving to answer this question thoroughly, we establish a neural checkpoint ranking benchmark (NeuCRaB) and study some intuitive ranking measures. These measures are generic, applying to the checkpoints of different output types without knowing how the checkpoints are pre-trained on which dataset. They also incur low computation cost, making them practically meaningful. Our results suggest that the linear separability of the features extracted by the checkpoints is a strong indicator of transferability. We also arrive at a new ranking measure, NLEEP, which gives rise to the best performance in the experiments.
****
SuperMix: Supervising the Mixing Data Augmentation
Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Nasser M. Nasrabadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13794-13803
This paper presents a supervised mixing augmentation method termed SuperMix, which exploits the salient regions within input images to construct mixed training samples. SuperMix is designed to obtain mixed images rich in visual features and

complying with realistic image priors. To enhance the efficiency of the algorithm, we develop a variant of the Newton iterative method, 65xfaster than gradient descent on this problem. We validate the effectiveness of SuperMix through extensive evaluations and ablation studies on two tasks of object classification and knowledge distillation. On the classification task, SuperMix provides comparable performance to the advanced augmentation methods, such as AutoAugment and RandAugment. In particular, combining SuperMix with RandAugment achieves 78.2% top-1 accuracy on ImageNet with ResNet50. On the distillation task, solely classifying images mixed using the teacher's knowledge achieves comparable performance to the state-of-the-art distillation methods. Furthermore, on average, incorporating mixed images into the distillation objective improves the performance by 3.4% and 3.1% on CIFAR-100 and ImageNet, respectively. The code is available at https://github.com/alldbi/SuperMix.
****
Informative and Consistent Correspondence Mining for Cross-Domain Weakly Supervised Object Detection
Luwei Hou, Yu Zhang, Kui Fu, Jia Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9929-9938
Cross-domain weakly supervised object detection aims to adapt object-level knowledge from a fully labeled source domain dataset (i.e. with object bounding boxes) to train object detectors for target domains that are weakly labeled (i.e. with image-level tags). Instead of domain-level distribution matching, as popularly adopted in the literature, we propose to learn pixel-wise cross-domain correspondences for more precise knowledge transfer. It is realized through a novel cross-domain co-attention scheme trained as region competition. In this scheme, the cross-domain correspondence module seeks for informative features on the target domain image, which after being warped to the source domain image, could best explain its annotations. Meanwhile, a collaborative mask generator competes to mask out the relevant target image region to make the remaining features uninformative. Such competitive learning strives to correlate the full foreground in cross-domain image pairs, revealing the accurate object extent in target domain. To alleviate the ambiguity of inter-domain correspondence learning, a domain-cycle consistency regularizer is futher proposed to leverage the more reliable intra-domain correspondence. The proposed approach achieves consistent improvements over existing approaches by a considerable margin, demonstrated by the experiments on various datasets.
****
Inception Convolution With Efficient Dilation Search
Jie Liu, Chuming Li, Feng Liang, Chen Lin, Ming Sun, Junjie Yan, Wanli Ouyang, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11486-11495
As a variant of standard convolution, a dilated convolution can control effective receptive fields and handle large scale variance of objects without introducing additional computational costs. To fully explore the potential of dilated convolution, we proposed a new type of dilated convolution (referred to as inception convolution), where the convolution operations have independent dilation patterns among different axes, channels and layers. To develop a practical method for learning complex inception convolution based on the data, a simple but effective search algorithm, referred to as efficient dilation optimization (EDO), is developed. Based on statistical optimization, the EDO method operates in a low-cost manner and is extremely fast when it is applied on large scale datasets. Empirical results validate that our method achieves consistent performance gains for image recognition, object detection, instance segmentation, human detection, and human pose estimation. For instance, by simply replacing the 3 x 3 standard convolution in the ResNet-50 backbone with inception convolution, we significantly improve the AP of Faster R-CNN from 36.4% to 39.2% on MS COCO.
****
Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy
Federico Paredes-Valles, Guido C. H. E. de Croon; Proceedings of the IEEE/CVF Co

nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3446-3455
Event cameras are novel vision sensors that sample, in an asynchronous fashion, brightness increments with low latency and high temporal resolution. The resulting streams of events are of high value by themselves, especially for high speed motion estimation. However, a growing body of work has also focused on the reconstruction of intensity frames from the events, as this allows bridging the gap with the existing literature on appearance- and frame-based computer vision. Recent work has mostly approached this problem using neural networks trained with synthetic, ground-truth data. In this work we approach, for the first time, the intensity reconstruction problem from a self-supervised learning perspective. Our method, which leverages the knowledge of the inner workings of event cameras, combines estimated optical flow and the event-based photometric constancy to train neural networks without the need for any ground-truth or synthetic data. Results across multiple datasets show that the performance of the proposed self-supervised approach is in line with the state-of-the-art. Additionally, we propose a novel, lightweight neural network for optical flow estimation that achieves high speed inference with only a minor drop in performance.
****
AdderSR: Towards Energy Efficient Image Super-Resolution
Dehua Song, Yunhe Wang, Hanting Chen, Chang Xu, Chunjing Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15648-15657
This paper studies the single image super-resolution problem using adder neural networks (AdderNets). Compared with convolutional neural networks, AdderNets utilize additions to calculate the output features thus avoid massive energy consumptions of conventional multiplications. However, it is very hard to directly inherit the existing success of AdderNets on large-scale image classification to the image super-resolution task due to the different calculation paradigm. Specifically, the adder operation cannot easily learn the identity mapping, which is essential for image processing tasks. In addition, the functionality of high-pass filters cannot be ensured by AdderNets. To this end, we thoroughly analyze the relationship between an adder operation and the identity mapping and insert shortcuts to enhance the performance of SR models using adder networks. Then, we develop a learnable power activation for adjusting the feature distribution and refining details. Experiments conducted on several benchmark models and datasets demonstrate that, our image super-resolution models using AdderNets can achieve comparable performance and visual quality to that of their CNN baselines with an about 2.5x reduction on the energy consumption. The codes are available at: https://github.com/huawei-noah/AdderNet.
****
Semi-Supervised Domain Adaptation Based on Dual-Level Domain Mixing for Semantic Segmentation
Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, Jianzhuang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11018-11027
Data-driven based approaches, in spite of great success in many tasks, have poor generalization when applied to unseen image domains, and require expensive cost of annotation especially for dense pixel prediction tasks such as semantic segmentation. Recently, both unsupervised domain adaptation (UDA) from large amounts of synthetic data and semi-supervised learning (SSL) with small set of labeled data have been studied to alleviate this issue. However, there is still a large gap on performance compared to their supervised counterparts. We focus on a more practical setting of semi-supervised domain adaptation (SSDA) where both a small set of labeled target data and large amounts of labeled source data are available. To address the task of SSDA, a novel framework based on dual-level domain mixing is proposed. The proposed framework consists of three stages. First, two kinds of data mixing methods are proposed to reduce domain gap in both region-level and sample-level respectively. We can obtain two complementary domain-mixed teachers based on dual-level mixed data from holistic and partial views respectively. Then, a student model is learned by distilling knowledge from these two tea

chers. Finally, pseudo labels of unlabeled data are generated in a self-training manner for another few rounds of teachers training. Extensive experimental results have demonstrated the effectiveness of our proposed framework on synthetic-to-real semantic segmentation benchmarks.
****

Connecting What To Say With Where To Look by Modeling Human Attention Traces
Zihang Meng, Licheng Yu, Ning Zhang, Tamara L. Berg, Babak Damavandi, Vikas Singh, Amy Bearman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12679-12688
We introduce a unified framework to jointly model images, text, and human attention traces. Our work is built on top of the recent Localized Narratives annotation framework, where each word of a given caption is paired with a mouse trace segment. We propose two novel tasks: (1) predict a trace given an image and caption (i.e., visual grounding), and (2) predict a caption and a trace given only an image. Learning the grounding of each word is challenging, due to noise in the human-provided traces and the presence of words that cannot be meaningfully visually grounded. We present a novel model architecture that is jointly trained on dual tasks (controlled trace generation and controlled caption generation). To evaluate the quality of the generated traces, we propose a local bipartite matching (LBM) distance metric which allows the comparison of two traces of different lengths. Extensive experiments show our model is robust to the imperfect training data and outperforms the baselines by a clear margin. Moreover, we demonstrate that our model pre-trained on the proposed tasks can be also beneficial to the downstream task of COCO's guided image captioning. Our code and project page are publicly available.
****

Shelf-Supervised Mesh Prediction in the Wild
Yufei Ye, Shubham Tulsiani, Abhinav Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8843-8852
We aim to infer 3D shape and pose of objects from a single image and propose a learning-based approach that can train from unstructured image collections, using only segmentation outputs from off-the-shelf recognition systems as supervisory signal (i.e. 'shelf-supervised'). We first infer a volumetric representation in a canonical frame, along with the camera pose for the input image. We enforce the representation to be geometrically consistent with both appearance and masks, and also that the synthesized novel views are indistinguishable from image collections. The coarse volumetric prediction is then converted to a mesh-based representation, which is further refined in the predicted camera frame. These two steps allow both shape-pose factorization from unannotated images and reconstruction of per-instance shape in finer details. We report performance on both synthetic and real-world datasets and demonstrate the scalability of our approach on 50 categories in the wild, an order of magnitude more classes than existing works.
****

Learning To Filter: Siamese Relation Network for Robust Tracking
Siyuan Cheng, Bineng Zhong, Guorong Li, Xin Liu, Zhenjun Tang, Xianxian Li, Jing Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4421-4431
Despite the great success of Siamese-based trackers, their performance under complicated scenarios is still not satisfying, especially when there are distractors. To this end, we propose a novel Siamese relation network, which introduces two efficient modules, i.e. Relation Detector (RD) and Refinement Module (RM). RD performs in a meta-learning way to obtain a learning ability to filter the distractors from the background while RM aims to effectively integrate the proposed RD into the Siamese framework to generate accurate tracking result.Moreover, to further improve the discriminability and robustness of the tracker, we introduce a contrastive training strategy that attempts not only to learn matching the same target but also to learn how to distinguish the different objects. Therefore, our tracker can achieve accurate tracking results when facing background clutters, fast motion, and occlusion. Experimental results on five popular benchmarks, including VOT2018, VOT2019, OTB100, LaSOT, and UAV123, show that the proposed me

thod is effective and can achieve state-of-the-art results. The code will be available at https://github.com/hqucv/siamrn
****
Ensembling With Deep Generative Views
Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, Richard Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14997-15007
Recent generative models can synthesize "views" of artificial images that mimic real-world variations, such as changes in color or pose, simply by learning from unlabeled image collections. Here, we investigate whether such views can be applied to real images to benefit downstream analysis tasks such as image classification. Using a pretrained generator, we first find the latent code corresponding to a given real input image. Applying perturbations to the code creates natural variations of the image, which can then be ensembled together at test-time. We use StyleGAN2 as the source of generative augmentations and investigate this setup on classification tasks involving facial attributes, cat faces, and cars. Critically, we find that several design decisions are required towards making this process work; the perturbation procedure, weighting between the augmentations and original image, and training the classifier on synthesized images can all impact the result. Currently, we find that while test-time ensembling with GAN-based augmentations can offer some small improvements, the remaining bottlenecks are the efficiency and accuracy of the GAN reconstructions, coupled with classifier sensitivities to artifacts in GAN-generated images.
****
Accurate Few-Shot Object Detection With Support-Query Mutual Guidance and Hybrid Loss
Lu Zhang, Shuigeng Zhou, Jihong Guan, Ji Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14424-14432
Most object detection methods require huge amounts of annotated data and can detect only the categories that appear in the training set. However, in reality acquiring massive annotated training data is both expensive and time-consuming. In this paper, we propose a novel two-stage detector for accurate few-shot object detection. In the first stage, we employ a support-query mutual guidance mechanism to generate more support-relevant proposals. Concretely, on the one hand, a query-guided support weighting module is developed for aggregating different supports to generate the support feature. On the other hand, a support-guided query enhancement module is designed by dynamic kernels. In the second stage, we score and filter proposals via multi-level feature comparison between each proposal and the aggregated support feature based on a distance metric learnt by an effective hybrid loss, which makes the embedding space of distance metric more discriminative. Extensive experiments on benchmark datasets show that our method substantially outperforms the existing methods and lifts the SOTA of FSOD task to a higher level.
****
Cascaded Prediction Network via Segment Tree for Temporal Video Grounding
Yang Zhao, Zhou Zhao, Zhu Zhang, Zhijie Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4197-4206
Temporal video grounding aims to localize the target segment which is semantically aligned with the given sentence in an untrimmed video. Existing methods can be divided into two main categories, including proposal-based approaches and proposal-free approaches. However, the former ones suffer from the extra cost of generating proposals and inflexibility in determining fine-grained boundaries, and the latter ones usually attempt to decide the start and end timestamps directly, which brings about much difficulty and inaccuracy. In this paper, we convert this task into a multi-step decision problem and propose a novel Cascaded Prediction Network (CPN) to generate the grounding result in a coarse-to-fine manner. Concretely, we first encode video and query into the same latent space and fuse them into integrated representations. Afterwards, we construct a segment-tree-based structure and make predictions via decision navigation and signal decomposition in a cascaded way. We evaluate our proposed method on three large-scale public

ly available benchmarks, namely ActivityNet Caption, Charades-STA and TACoS, whe re our CPN surpasses the performance of the state-of-the-art methods.
****
Posterior Promoted GAN With Distribution Discriminator for Unsupervised Image Sy nthesis
Xianchao Zhang, Ziyang Cheng, Xiaotong Zhang, Han Liu; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6519-6528
Sufficient real information in generator is a critical point for the generation ability of GAN. However, GAN and its variants suffer from lack of this point, re sulting in brittle training processes. In this paper, we propose a novel variant of GAN, Posterior Promoted GAN (P2GAN), which promotes generator with the real information in the posterior distribution produced by discriminator. In our fram ework, different from other variants of GAN, the discriminator maps images to a multivariate Gaussian distribution and extracts real information. The generator employs the real information by AdaIN and a latent code regularizer. Besides, re parameterization trick and pretraining is applied to guarantee a stable training process in practice. The convergence of P2GAN is theoretically proved. Experime ntal results on typical high-dimensional multi-modal datasets demonstrate that P 2GAN has achieved comparable results with the state-of-the-art variants of GAN o n unsupervised image synthesis.
****
Toward Accurate and Realistic Outfits Visualization With Attention to Details
Kedan Li, Min Jin Chong, Jeffrey Zhang, Jingen Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15546-15 555
Virtual try-on methods aim to generate images of fashion models wearing arbitrar y combinations of garments. This is a challenging task because the generated ima ge must appear realistic and accurately display the interaction between garments . Prior works produce images that are filled with artifacts and fail to capture important visual details necessary for commercial applications. We propose Outfi t Visualization Net (OVNet) to capture these important details (e.g. buttons, sh ading, textures, realistic hemlines, and interactions between garments) and prod uce high quality multiple-garment virtual try-on images. OVNet consists of 1) a semantic layout generator and 2) an image generation pipeline using multiple coo rdinated warps. We train the warper to output multiple warps using a cascade los s, which refines each successive warp to focus on poorly generated regions of a previous warp and yields consistent improvements in detail. In addition, we intr oduce a method for matching outfits with the most suitable model and produce sig nificant improvements for both our and other previous try-on methods. Through qu antitative and qualitative analysis, we demonstrate our method generates substan tially higher-quality studio images compared to prior works for multi-garment ou tfits. An interactive interface powered by this method has been deployed on fash ion e-commerce websites and received overwhelmingly positive feedback.
****
Delving Deep Into Many-to-Many Attention for Few-Shot Video Object Segmentation
Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14040-14049
This paper tackles the task of Few-Shot Video Object Segmentation (FSVOS), i.e., segmenting objects in the query videos with certain class specified in a few la beled support images. The key is to model the relationship between the query vid eos and the support images for propagating the object information. This is a man y-to-many problem and often relies on full-rank attention, which is computationa lly intensive. In this paper, we propose a novel Domain Agent Network (DAN), bre aking down the full-rank attention into two smaller ones. We consider one single frame of the query video as the domain agent, bridging between the support imag es and the query video. Our DAN allows a linear space and time complexity as opp osed to the original quadratic form with no loss of performance. In addition, we introduce a learning strategy by combining meta-learning with online learning t

o further improve the segmentation accuracy. We build a FSVOS benchmark on the Y outube-VIS dataset and conduct experiments to demonstrate that our method outper forms baselines on both computational cost and accuracy, achieving the state-of-the-art performance.
****
MongeNet: Efficient Sampler for Geometric Deep Learning
Leo Lebrat, Rodrigo Santa Cruz, Clinton Fookes, Olivier Salvado; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16664-16673
Recent advances in geometric deep-learning introduce complex computational chall enges for evaluating the distance between meshes. From a mesh model, point cloud s are necessary along with a robust distance metric to assess surface quality or as part of the loss function for training models. Current methods often rely on a uniform random mesh discretization, which yields irregular sampling and noisy distance estimation. In this paper we introduce MongeNet, a fast and optimal tr ansport based sampler that allows for an accurate discretization of a mesh with better approximation properties. We compare our method to the ubiquitous random uniform sampling and show that the approximation error is almost half with a ver y small computational overhead.
****
Gated Spatio-Temporal Attention-Guided Video Deblurring
Maitreya Suin, A. N. Rajagopalan; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2021, pp. 7802-7811
Video deblurring remains a challenging task due to the complexity of spatially a nd temporally varying blur. Most of the existing works depend on implicit or exp licit alignment for temporal information fusion, which either increases the comp utational cost or results in suboptimal performance due to misalignment. In this work, we investigate two key factors responsible for deblurring quality: how to fuse spatio-temporal information and from where to collect it. We propose a fac torized gated spatio-temporal attention module to perform non-local operations a cross space and time to fully utilize the available information without dependin g on alignment. First, we perform spatial aggregation followed by a temporal agg regation step. Next, we adaptively distribute the global spatio-temporal informa tion to each pixel. It shows superior performance compared to existing non-local fusion techniques while being considerably more efficient. To complement the at tention module, we propose a reinforcement learning-based framework for selectin g keyframes from the neighborhood with the most complementary and useful informa tion. Moreover, our adaptive approach can increase or decrease the frame usage a t inference time, depending on the user's need. Extensive experiments on multipl e datasets demonstrate the superiority of our method.
****
Learning Multi-Scale Photo Exposure Correction
Mahmoud Afifi, Konstantinos G. Derpanis, Bjorn Ommer, Michael S. Brown; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9157-9167
Capturing photographs with wrong exposures remains a major source of errors in c amera-based imaging. Exposure problems are categorized as either: (i) overexpose d, where the camera exposure was too long, resulting in bright and washed-out im age regions, or (ii) underexposed, where the exposure was too short, resulting i n dark regions. Both under- and overexposure greatly reduce the contrast and vis ual appeal of an image. Prior work mainly focuses on underexposed images or gene ral image enhancement. In contrast, our proposed method targets both over- and u nder-exposure errors in photographs. We formulate the exposure correction proble m as two main sub-problems: (i) color enhancement and (ii) detail enhancement. A ccordingly, we propose a coarse-to-fine deep neural network (DNN) model, trainab le in an end-to-end manner, that addresses each sub-problem separately. A key as pect of our solution is a new dataset of over 24,000 images exhibiting the broad est range of exposure values to date with a corresponding properly exposed image . Our method achieves results on par with existing state-of-the-art methods on u nderexposed images and yields significant improvements for images suffering from

overexposure errors.
****
Learning Semantic Person Image Generation by Region-Adaptive Normalization
Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, Wangmeng Zuo
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2021, pp. 10806-10815
Human pose transfer has received great attention due to its wide applications, y
et is still a challenging task that is not well solved. Recent works have achiev
ed great success to transfer the person image from the source to the target pose
. However, most of them cannot well capture the semantic appearance, resulting i
n inconsistent and less realistic textures on the reconstructed results. To addr
ess this issue, we propose a new two-stage framework to handle the pose and appe
arance translation. In the first stage, we predict the target semantic parsing m
aps to eliminate the difficulties of pose transfer and further benefit the latte
r translation of per-region appearance style. In the second one, with the predic
ted target semantic maps, we suggest a new person image generation method by inc
orporating the region-adaptive normalization, in which it takes the per-region s
tyles to guide the target appearance generation. Extensive experiments show that
 our proposed SPGNet can generate more semantic, consistent, and photo-realistic
 results and perform favorably against the state of the art methods in terms of
quantitative and qualitative evaluation. The source code and model are available
 at https://github.com/cszy98/SPGNet.git.
****
Rethinking Class Relations: Absolute-Relative Supervised and Unsupervised Few-Sh
ot Learning
Hongguang Zhang, Piotr Koniusz, Songlei Jian, Hongdong Li, Philip H. S. Torr; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2021, pp. 9432-9441
The majority of existing few-shot learning methods describe image relations with
 binary labels. However, such binary relations are insufficient to teach the net
work complicated real-world relations, due to the lack of decision smoothness. F
urthermore, current few-shot learning models capture only the similarity via rel
ation labels, but they are not exposed to class concepts associated with objects
, which is likely detrimental to the classification performance due to underutil
ization of the available class labels. For instance, children learn the concept
of tiger from a few of actual examples as well as from comparisons of tiger to o
ther animals. Thus, we hypothesize that both similarity and class concept learni
ng must be occurring simultaneously. With these observations at hand, we study t
he fundamental problem of simplistic class modeling in current few-shot learning
 methods. We rethink the relations between class concepts, and propose a novel A
bsolute-relative Learning paradigm to fully take advantage of label information
to refine the image an relation representations in both supervised and unsupervi
sed scenarios. Our proposed paradigm improves the performance of several state-o
f-the-art models on publicly available datasets.
****
Divergence Optimization for Noisy Universal Domain Adaptation
Qing Yu, Atsushi Hashimoto, Yoshitaka Ushiku; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2515-2524
Universal domain adaptation (UniDA) has been proposed to transfer knowledge lear
ned from a label-rich source domain to a label-scarce target domain without any
constraints on the label sets. In practice, however, it is difficult to obtain a
 large amount of perfectly clean labeled data in a source domain with limited re
sources. Existing UniDA methods rely on source samples with correct annotations,
 which greatly limits their application in the real world. Hence, we consider a
new realistic setting called Noisy UniDA, in which classifiers are trained with
noisy labeled data from the source domain and unlabeled data with an unknown cla
ss distribution from the target domain. This paper introduces a two-head convolu
tional neural network framework to solve all problems simultaneously. Our networ
k consists of one common feature generator and two classifiers with different de
cision boundaries. By optimizing the divergence between the two classifiers' out

puts, we can detect noisy source samples, find "unknown" classes in the target domain, and align the distribution of the source and target domains. In an extensive evaluation of different domain adaptation settings, the proposed method outperformed existing methods by a large margin in most settings.
****
Learning Dynamic Alignment via Meta-Filter for Few-Shot Learning
Chengming Xu, Yanwei Fu, Chen Liu, Chengjie Wang, Jilin Li, Feiyue Huang, Li Zhang, Xiangyang Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5182-5191
Few-shot learning (FSL), which aims to recognise new classes by adapting the learned knowledge with extremely limited few-shot (support) examples, remains an important open problem in computer vision. Most of the existing methods for feature alignment in few-shot learning only consider image-level or spatial-level alignment while omitting the channel disparity. Our insight is that these methods would lead to poor adaptation with redundant matching, and leveraging channel-wise adjustment is the key to well adapting the learned knowledge to new classes. Therefore, in this paper, we propose to learn a dynamic alignment, which can effectively highlight both query regions and channels according to different local support information. Specifically, this is achieved by first dynamically sampling the neighbourhood of the feature position conditioned on the input few shot, based on which we further predict a both position-dependent and channel-dependent Dynamic Meta-filter. The filter is used to align the query feature with position-specific and channel-specific knowledge. Moreover, we adopt Neural Ordinary Differential Equation (ODE) to enable a more accurate control of the alignment. In such a sense our model is able to better capture fine-grained semantic context of the few-shot example and thus facilitates dynamical knowledge adaptation for few-shot learning. The resulting framework establishes the new state-of-the-arts on major few-shot visual recognition benchmarks, including miniImageNet and tieredImageNet.
****
Unsupervised Learning of 3D Object Categories From Videos in the Wild
Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, David Novotny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4700-4709
Recently, numerous works have attempted to learn 3D reconstructors of textured 3D models of visual categories given a training set of annotated static images of objects. In this paper, we seek to decrease the amount of needed supervision by leveraging a collection of object-centric videos captured in-the-wild without requiring any manual 3D annotations. Since existing category-centric datasets are insufficient for this problem, we contribute with a large-scale crowd-sourced dataset of object-centric videos suitable for this task. We further propose a novel method that learns via differentiable rendering of a predicted implicit surface of the scene. Here, inspired by classic multi-view stereo methods, our key technical contribution is a novel warp-conditioned implicit shape function, which is robust to the noise in the SfM video reconstructions that supervise our learning. Our evaluation demonstrates performance improvements over several deep monocular reconstruction baselines on 2 existing benchmarks and on our novel dataset.
****
Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing
Yu Wu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1326-1335
We investigate the weakly-supervised audio-visual video parsing task, which aims to parse a video into temporal event segments and predict the audible or visible event categories. The task is challenging since there only exist video-level event labels for training, without indicating the temporal boundaries and modalities. Previous works take the overall event labels to supervise both audio and visual model predictions. However, we argue that such overall labels harm the model training due to the audio-visual asynchrony. For example, commentators speak in a basketball video, but we cannot visually find the speakers. In this paper, w

e tackle this issue by leveraging the cross-modal correspondence of audio and visual signals. We generate reliable event labels individually for each modality by swapping audio and visual tracks with other unrelated videos. If the original visual/audio data contain event clues, the event prediction from the newly assembled data would still be highly confident. In this way, we could protect our models from being misled by ambiguous event labels. In addition, we propose the cross-modal audio-visual contrastive learning to induce temporal difference on attention models within videos, i.e., urging the model to pick the current temporal segment from all context candidates. Experiments show we outperform state-of-the-art methods by a large margin.
****
Dogfight: Detecting Drones From Drones Videos
Muhammad Waseem Ashraf, Waqas Sultani, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7067-7076
As airborne vehicles are becoming more autonomous and ubiquitous, it has become vital to develop the capability to detect the objects in their surroundings. This paper attempts to address the problem of drones detection from other flying drones. The erratic movement of the source and target drones, small size, arbitrary shape, large intensity variations, and occlusion make this problem quite challenging. In this scenario, region-proposal based methods are not able to capture sufficient discriminative foreground-background information. Also, due to the extremely small size and complex motion of the source and target drones, feature aggregation based methods are unable to perform well. To handle this, instead of using region-proposal based methods, we propose to use a two-stage segmentation-based approach employing spatio-temporal attention cues. During the first stage, given the overlapping frame regions, detailed contextual information is captured over convolution feature maps using pyramid pooling. After that pixel and channel-wise attention is enforced on the feature maps to ensure accurate drone localization. In the second stage, first stage detections are verified and new probable drone locations are explored. To discover new drone locations, motion boundaries are used. This is followed by tracking candidate drone detections for a few frames, cuboid formation, extraction of the 3D convolution feature map, and drones detection within each cuboid. The proposed approach is evaluated on two publicly available drone detection datasets and outperforms over several competitive baselines.
****
PAUL: Procrustean Autoencoder for Unsupervised Lifting
Chaoyang Wang, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 434-443
Recent success in casting Non-rigid Structure from Motion (NRSfM) as an unsupervised deep learning problem has raised fundamental questions about what novelty in NRSfM prior could the deep learning offer. In this paper we advocate for a 3D deep auto-encoder framework to be used explicitly as the NRSfM prior. The framework is unique as: (i) it learns the 3D auto-encoder weights solely from 2D projected measurements, and (ii) it is Procrustean in that it jointly resolves the unknown rigid pose for each shape instance. We refer to this architecture as a Procustean Autoencoder for Unsupervised Lifting (PAUL), and demonstrate state-of-the-art performance across a number of benchmarks in comparison to recent innovations such as Deep NRSfM and C3PDO.
****
Group Collaborative Learning for Co-Salient Object Detection
Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12288-12298
We present a novel group collaborative learning framework (GCNet) capable of detecting co-salient objects in real time (16ms), by simultaneously mining consensus representations at group level based on the two necessary criteria: 1) intra-group compactness to better formulate the consistency among co-salient objects by capturing their inherent shared attributes using our novel group affinity modul

e; 2) inter-group separability to effectively suppress the influence of noisy objects on the output by introducing our new group collaborating module conditioning the inconsistent consensus. To learn a better embedding space without extra computational overhead, we explicitly employ auxiliary classification supervision. Extensive experiments on three challenging benchmarks, i.e., CoCA, CoSOD3k, and Cosal2015, demonstrate that our simple GCNet outperforms 10 cutting-edge models and achieves the new state-of-the-art. We demonstrate this paper's new technical contributions on a number of important downstream computer vision applications including content aware co-segmentation, co-localization based automatic thumbnails, etc. Our research code with two applications will be released.
****
RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening
Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11580-11590
Enhancing the generalization capability of deep neural networks to unseen domains is crucial for safety-critical applications in the real world such as autonomous driving. To address this issue, this paper proposes a novel instance selective whitening loss to improve the robustness of the segmentation networks for unseen domains. Our approach disentangles the domain-specific style and domain-invariant content encoded in higher-order statistics (i.e., feature covariance) of the feature representations and selectively removes only the style information causing domain shift. As shown in Fig. 1, our method provides reasonable predictions for (a) low-illuminated, (b) rainy, and (c) unseen structures. These types of images are not included in the training dataset, where the baseline shows a significant performance drop, contrary to ours. Being simple yet effective, our approach improves the robustness of various backbone networks without additional computational cost. We conduct extensive experiments in urban-scene segmentation and show the superiority of our approach to existing work.
****
Monocular Real-Time Full Body Capture With Inter-Part Correlations
Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4811-4822
We present the first method for real-time full body capture that estimates shape and motion of body and hands together with a dynamic 3D face model from a single color image. Our approach uses a new neural network architecture that exploits correlations between body and hands at high computational efficiency. Unlike previous works, our approach is jointly trained on multiple datasets focusing on hand, body or face separately, without requiring data where all the parts are annotated at the same time, which is much more difficult to create at sufficient variety. The possibility of such multi-dataset training enables superior generalization ability. In contrast to earlier monocular full body methods, our approach captures more expressive 3D face geometry and color by estimating the shape, expression, albedo and illumination parameters of a statistical face model. Our method achieves competitive accuracy on public benchmarks, while being significantly faster and providing more complete face reconstructions.
****
Pre-Trained Image Processing Transformer
Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, Wen Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12299-12310
As the computing power of modern hardware is increasing strongly, pre-trained deep learning models (e.g., BERT, GPT-3) learned on large-scale datasets have shown their effectiveness over conventional methods. The big progress is mainly contributed to the representation ability of transformer and its variant architectures. In this paper, we study the low-level computer vision task (e.g., denoising, super-resolution and deraining) and develop a new pre-trained model, namely, image processing transformer (IPT). To maximally excavate the capability of transf

ormer, we present to utilize the well-known ImageNet benchmark for generating a large amount of corrupted image pairs. The IPT model is trained on these images with multi-heads and multi-tails. In addition, the constructive learning is introduced for well adapting to different image processing tasks. The pre-trained model can therefore efficiently employed on desired task after fine-tuning. With only one pre-trained model, IPT outperforms the current state-of-the-art methods on various low-level benchmarks. Code is available at https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/IPT
****
Robust and Accurate Object Detection via Adversarial Learning
Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, Boqing Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16622-16631
Data augmentation has become a de facto component for training high-performance deep image classifiers, but its potential is under-explored for object detection. Noting that most state-of-the-art object detectors benefit from fine-tuning a pre-trained classifier, we first study how the classifiers' gains from various data augmentations transfer to object detection. The results are discouraging; the gains diminish after fine-tuning in terms of either accuracy or robustness. This work instead augments the fine-tuning stage for object detectors by exploring adversarial examples, which can be viewed as a model-dependent data augmentation. Our method dynamically selects the stronger adversarial images sourced from a detector's classification and localization branches and evolves with the detector to ensure the augmentation policy stays current and relevant. This model-dependent augmentation generalizes to different object detectors better than AutoAugment, a model-agnostic augmentation policy searched based on one particular detector. Our approach boosts the performance of state-of-the-art EfficientDets by +1.1 mAP on the COCO object detection benchmark. It also improves the detectors' robustness against natural distortions by +3.8 mAP and against domain shift by +1.3 mAP.
****
Faster Meta Update Strategy for Noise-Robust Deep Learning
Youjiang Xu, Linchao Zhu, Lu Jiang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 144-153
It has been shown that deep neural networks are prone to overfitting on biased training data. Towards addressing this issue, meta-learning employs a meta model for correcting the training bias. Despite the promising performances, super slow training is currently the bottleneck in the meta learning approaches. In this paper, we introduce a novel Faster Meta Update Strategy (FaMUS) to replace the most expensive step in the meta gradient computation with a faster layer-wise approximation. We empirically find that FaMUS yields not only a reasonably accurate but also a low-variance approximation of the meta gradient. We conduct extensive experiments to verify the proposed method on two tasks. We show our method is able to save two-thirds of the training time while still maintaining the comparable or achieving even better generalization performance. In particular, our method achieves the state-of-the-art performance on both synthetic and realistic noisy labels, and obtains promising performance on long-tailed recognition on standard benchmarks. Code are released at https://github.com/youjiangxu/FaMUS.
****
ContactOpt: Optimizing Contact To Improve Grasps
Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, Charles C. Kemp; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1471-1481
Physical contact between hands and objects plays a critical role in human grasps. We show that optimizing the pose of a hand to achieve expected contact with an object can improve hand poses inferred via image-based methods. Given a hand mesh and an object mesh, a deep model trained on ground truth contact data infers desirable contact across the surfaces of the meshes. Then, ContactOpt efficiently optimizes the pose of the hand to achieve desirable contact using a differentiable contact model. Notably, our contact model encourages mesh interpenetration

to approximate deformable soft tissue in the hand. In our evaluations, our methods result in grasps that better match ground truth contact, have lower kinematic error, and are significantly preferred by human participants. Code and models are available online.
****
Panoptic-PolarNet: Proposal-Free LiDAR Point Cloud Panoptic Segmentation
Zixiang Zhou, Yang Zhang, Hassan Foroosh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13194-13203
Panoptic segmentation presents a new challenge in exploiting the merits of both detection and segmentation, with the aim of unifying instance segmentation and semantic segmentation in a single framework. However, an efficient solution for panoptic segmentation in the emerging domain of LiDAR point cloud is still an open research problem and is very much under-explored. In this paper, we present a fast and robust LiDAR point cloud panoptic segmentation framework, referred to as Panoptic-PolarNet. We learn both semantic segmentation and class-agnostic instance clustering in a single inference network using a polar Bird's Eye View (BEV) representation, enabling us to circumvent the issue of occlusion among instances in urban street scenes. To improve our network's learnability, we also propose an adapted instance augmentation technique and a novel adversarial point cloud pruning method. Our experiments show that Panoptic-PolarNet outperforms the baseline methods on SemanticKITTI and nuScenes datasets with an almost real-time inference speed. Panoptic-PolarNet achieved 54.1% PQ in the public SemanticKITTI panoptic segmentation leaderboard and leading performance for the validation set of nuScenes.
****
Source-Free Domain Adaptation for Semantic Segmentation
Yuang Liu, Wei Zhang, Jun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1215-1224
Unsupervised Domain Adaptation (UDA) can tackle the challenge that convolutional neural network (CNN)-based approaches for semantic segmentation heavily rely on the pixel-level annotated data, which is labor-intensive. However, existing UDA approaches in this regard inevitably require the full access to source datasets to reduce the gap between the source and target domains during model adaptation, which are impractical in the real scenarios where the source datasets are private, and thus cannot be released along with the well-trained source models. To cope with this issue, we propose a source-free domain adaptation framework for semantic segmentation, namely SFDA, in which only a well-trained source model and an unlabeled target domain dataset are available for adaptation. SFDA not only enables to recover and preserve the source domain knowledge from the source model via knowledge transfer during model adaptation, but also distills valuable information from the target domain for self-supervised learning. The pixel- and patch-level optimization objectives tailored for semantic segmentation are seamlessly integrated in the framework. The extensive experimental results on numerous benchmark datasets highlight the effectiveness of our framework against the existing UDA approaches relying on source data.
****
Adaptive Weighted Discriminator for Training Generative Adversarial Networks
Vasily Zadorozhnyy, Qiang Cheng, Qiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4781-4790
Generative adversarial network (GAN) has become one of the most important neural network models for classical unsupervised machine learning. A variety of discriminator loss functions have been developed to train GAN's discriminators and they all have a common structure: a sum of real and fake losses that only depends on the actual and generated data respectively. One challenge associated with an equally weighted sum of two losses is that the training may benefit one loss but harm the other, which we show causes instability and mode collapse. In this paper, we introduce a new family of discriminator loss functions that adopts a weighted sum of real and fake parts, which we call adaptive weighted loss functions or aw-loss functions. Using the gradients of the real and fake parts of the loss, we can adaptively choose weights to train a discriminator in the direction that

benefits the GAN's stability. Our method can be potentially applied to any disc
riminator model with a loss that is a sum of the real and fake parts. Experiment
s validated the effectiveness of our loss functions on unconditional and conditi
onal image generation tasks, improving the baseline results by a significant mar
gin on CIFAR-10, STL-10, and CIFAR-100 datasets in Inception Scores (IS) and Fre
chet Inception Distance (FID) metrics.
****
Depth From Camera Motion and Object Detection
Brent A. Griffin, Jason J. Corso; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2021, pp. 1397-1406
This paper addresses the problem of learning to estimate the depth of detected o
bjects given some measurement of camera motion (e.g., from robot kinematics or v
ehicle odometry). We achieve this by 1) designing a recurrent neural network (DB
ox) that estimates the depth of objects using a generalized representation of bo
unding boxes and uncalibrated camera movement and 2) introducing the Object Dept
h via Motion and Detection Dataset (ODMD). ODMD training data are extensible and
configurable, and the ODMD benchmark includes 21,600 examples across four valid
ation and test sets. These sets include mobile robot experiments using an end-ef
fector camera to locate objects from the YCB dataset and examples with perturbat
ions added to camera motion or bounding box data. In addition to the ODMD benchm
ark, we evaluate DBox in other monocular application domains, achieving state-of
-the-art results on existing driving and robotics benchmarks and estimating the
depth of objects using a camera phone.
****
PPR10K: A Large-Scale Portrait Photo Retouching Dataset With Human-Region Mask a
nd Group-Level Consistency
Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, Lei Zhang; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
653-661
Different from general photo retouching tasks, portrait photo retouching (PPR),
which aims to enhance the visual quality of a collection of flat-looking portrai
t photos, has its special and practical requirements such as human-region priori
ty (HRP) and group-level consistency (GLC). HRP requires that more attention sho
uld be paid to human regions, while GLC requires that a group of portrait photos
should be retouched to a consistent tone. Models trained on existing general ph
oto retouching datasets, however, can hardly meet these requirements of PPR. To
facilitate the research on this high-frequency task, we construct a large-scale
PPR dataset, namely PPR10K, which is the first of its kind to our best knowledge
. PPR10K contains 1, 681 groups and 11, 161 high-quality raw portrait photos in
total. High-resolution segmentation masks of human regions are provided. Each ra
w photo is retouched by three experts, while they elaborately adjust each group
of photos to have consistent tones. We define a set of objective measures to eva
luate the performance of PPR and propose strategies to learn PPR models with goo
d HRP and GLC performance. The constructed PPR10K dataset provides a good benchm
ark for studying automatic PPR methods, and experiments demonstrate that the pro
posed learning strategies are effective to improve the retouching performance. D
atasets and codes are available: https://github.com/csjliang/PPR10K.
****
Transformation Driven Visual Reasoning
Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, Xueqi Cheng; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
6903-6912
This paper defines a new visual reasoning paradigm by introducing an important f
actor, i.e. transformation. The motivation comes from the fact that most existin
g visual reasoning tasks, such as CLEVR in VQA, are solely defined to test how w
ell the machine understands the concepts and relations within static settings, l
ike one image. We argue that this kind of state driven visual reasoning approach
has limitations in reflecting whether the machine has the ability to infer the
dynamics between different states, which has been shown as important as state-le
vel reasoning for human cognition in Piaget's theory. To tackle this problem, we

propose a novel transformation driven visual reasoning task. Given both the initial and final states, the target is to infer the corresponding single-step or multi-step transformation, represented as a triplet (object, attribute, value) or a sequence of triplets, respectively. Following this definition, a new dataset namely TRANCE is constructed on the basis of CLEVR, including three levels of settings, i.e. Basic (single-step transformation), Event (multi-step transformation), and View (multi-step transformation with variant views). Experimental results show that the state-of-the-art visual reasoning models perform well on Basic, but are still far from human-level intelligence on Event and View. We believe the proposed new paradigm will boost the development of machine visual reasoning. More advanced methods and real data need to be investigated in this direction. The resource of TVR is available at https://hongxin2019.github.io/TVR.
****
Sparse R-CNN: End-to-End Object Detection With Learnable Proposals
Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14454-14463
We present Sparse R-CNN, a purely sparse method for object detection in images. Existing works on object detection heavily rely on dense object candidates, such as k anchor boxes pre-defined on all grids of image feature map of size HxW. In our method, however, a fixed sparse set of learned object proposals, total length of N, are provided to object recognition head to perform classification and location. By eliminating HWk (up to hundreds of thousands) hand-designed object candidates to N (e.g. 100) learnable proposals, Sparse R-CNN completely avoids all efforts related to object candidates design and many-to-one label assignment. More importantly, final predictions are directly output without non-maximum suppression post-procedure. Sparse R-CNN demonstrates accuracy, run-time and training convergence performance on par with the well-established detector baselines on the challenging COCO dataset, e.g., achieving 45.0 AP in standard 3x training schedule and running at 22 fps using ResNet-50 FPN model. We hope our work could inspire re-thinking the convention of dense prior in object detectors. The code is available at: https://github.com/PeizeSun/SparseR-CNN.
****
Plan2Scene: Converting Floorplans to 3D Scenes
Madhawa Vidanapathirana, Qirui Wu, Yasutaka Furukawa, Angel X. Chang, Manolis Savva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10733-10742
We address the task of converting a floorplan and a set of associated photos of a residence into a textured 3D mesh model, a task which we call Plan2Scene. Our system 1) lifts a floorplan image to a 3D mesh model; 2) synthesizes surface textures based on the input photos; and 3) infers textures for unobserved surfaces using a graph neural network architecture. To train and evaluate our system we create indoor surface texture datasets, and augment a dataset of floorplans and photos from prior work with rectified surface crops and additional annotations. Our approach handles the challenge of producing tileable textures for dominant surfaces such as floors, walls, and ceilings from a sparse set of unaligned photos that only partially cover the residence. Qualitative and quantitative evaluations show that our system produces realistic 3D interior models, outperforming baseline approaches on a suite of texture quality metrics and as measured by a holistic user study.
****
Towards Semantic Segmentation of Urban-Scale 3D Point Clouds: A Dataset, Benchmarks and Challenges
Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, Andrew Markham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4977-4987
An essential prerequisite for unleashing the potential of supervised deep learning algorithms in the area of 3D scene understanding is the availability of large-scale and richly annotated datasets. However, publicly available datasets are e

ither in relatively small spatial scales or have limited semantic annotations du
e to the expensive cost of data acquisition and data annotation, which severely
limits the development of fine-grained semantic understanding in the context of
3D point clouds. In this paper, we present an urban-scale photogrammetric point
cloud dataset with nearly three billion richly annotated points, which is three
times the number of labeled points than the existing largest photogrammetric poi
nt cloud dataset. Our dataset consists of large areas from three UK cities, cove
ring about 7.6 km^2 of the city landscape. In the dataset, each 3D point is labe
led as one of 13 semantic classes. We extensively evaluate the performance of st
ate-of-the-art algorithms on our dataset and provide a comprehensive analysis of
 the results. In particular, we identify several key challenges towards urban-sc
ale point cloud understanding. The dataset is available at https://github.com/Qi
ngyongHu/SensatUrban.
****

Towards Open World Object Detection
K J Joseph, Salman Khan, Fahad Shahbaz Khan, Vineeth N Balasubramanian; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2021, pp. 5830-5840
Humans have a natural instinct to identify unknown object instances in their env
ironments. The intrinsic curiosity about these unknown instances aids in learnin
g about them, when the corresponding knowledge is eventually available. This mot
ivates us to propose a novel computer vision problem called: `Open World Object
Detection', where a model is tasked to: 1) identify objects that have not been i
ntroduced to it as `unknown', without explicit supervision to do so, and 2) incr
ementally learn these identified unknown categories without forgetting previousl
y learned classes, when the corresponding labels are progressively received. We
formulate the problem, introduce a strong evaluation protocol and provide a nove
l solution, which we call OREO: Open World Object Detector, based on contrastive
 clustering and energy based unknown identification. Our experimental evaluation
 and ablation studies analyse the efficacy of OREO in achieving Open World objec
tives. As an interesting by-product, we find that identifying and characterising
 unknown instances helps to reduce confusion in an incremental object detection
setting, where we achieve state-of-the-art performance, with no extra methodolog
ical effort. We hope that our work will attract further research into this newly
 identified, yet crucial research direction.
****

Conditional Bures Metric for Domain Adaptation
You-Wei Luo, Chuan-Xian Ren; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2021, pp. 13989-13998
As a vital problem in classification-oriented transfer, unsupervised domain adap
tation (UDA) has attracted widespread attention in recent years. Previous UDA me
thods assume the marginal distributions of different domains are shifted while i
gnoring the discriminant information in the label distributions. This leads to c
lassification performance degeneration in real applications. In this work, we fo
cus on the conditional distribution shift problem which is of great concern to c
urrent conditional invariant models. We aim to seek a kernel covariance embeddin
g for conditional distribution which remains yet unexplored. Theoretically, we p
ropose the Conditional Kernel Bures (CKB) metric for characterizing conditional
distribution discrepancy, and derive an empirical estimation for the CKB metric
without introducing the implicit kernel feature map. It provides an interpretabl
e approach to understand the knowledge transfer mechanism. The established consi
stency theory of the empirical estimation provides a theoretical guarantee for c
onvergence. A conditional distribution matching network is proposed to learn the
 conditional invariant and discriminative features for UDA. Extensive experiment
s and analysis show the superiority of our proposed model.
****

DatasetGAN: Efficient Labeled Data Factory With Minimal Human Effort
Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Bar
riuso, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10145-10155

We introduce DatasetGAN: an automatic procedure to generate massive datasets of high-quality semantically segmented images requiring minimal human effort. Curre nt deep networks are extremely data-hungry, benefiting from training on large-sc ale datasets, which are time-consuming to annotate. Our method relies on the pow er of recent GANs to generate realistic images. We show how the GAN latent code can be decoded to produce a semantic segmentation of the image. Training the dec oder only needs a few labeled examples to generalize to the rest of the latent s pace, resulting in an infinite annotated dataset generator! These generated data sets can then be used for training any computer vision architecture just as real datasets are. As only a few images need to be manually segmented, it becomes po ssible to annotate images in extreme detail and generate datasets with rich obje ct and part segmentations. To showcase the power of our approach, we generated d atasets for 7 image segmentation tasks which include pixel-level labels for 34 h uman face parts, and 32 car parts. Our approach outperforms all semi-supervised baselines significantly and is on par with fully supervised methods using labor-intensive annotations.
****
Repurposing GANs for One-Shot Semantic Part Segmentation
Nontawat Tritrong, Pitchaporn Rewatbowornwong, Supasorn Suwajanakorn; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4475-4485
While GANs have shown success in realistic image generation, the idea of using G ANs for other tasks unrelated to synthesis is underexplored. Do GANs learn meani ngful structural parts of objects during their attempt to reproduce those object s? In this work, we test this hypothesis and propose a simple and effective appr oach based on GANs for semantic part segmentation that requires as few as one la bel example along with an unlabeled dataset. Our key idea is to leverage a train ed GAN to extract a pixel-wise representation from the input image and use it as feature vectors for a segmentation network. Our experiments demonstrate that th is GAN-derived representation is "readily discriminative" and produces surprisin gly good results that are comparable to those from supervised baselines trained with significantly more labels. We believe this novel repurposing of GANs underl ies a new class of unsupervised representation learning, which can generalize to many other tasks.
****
Semi-Supervised 3D Hand-Object Poses Estimation With Interactions in Time
Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, Xiaolong Wang; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14687-14697
Estimating 3D hand and object pose from a single image is an extremely challengi ng problem: hands and objects are often self-occluded during interactions, and t he 3D annotations are scarce as even humans cannot directly label the ground-tru ths from a single image perfectly. To tackle these challenges, we propose a unif ied framework for estimating the 3D hand and object poses with semi-supervised l earning. We build a joint learning framework where we perform explicit contextua l reasoning between hand and object representations. Going beyond limited 3D ann otations in a single image, we leverage the spatial-temporal consistency in larg e-scale hand-object videos as a constraint for generating pseudo labels in semi-supervised learning. Our method not only improves hand pose estimation in challe nging real-world dataset, but also substantially improve the object pose which h as fewer ground-truths per instance. By training with large-scale diverse videos , our model also generalizes better across multiple out-of-domain datasets. Proj ect page and code: https://stevenlsw.github.io/Semi-Hand-Object
****
Cyclic Co-Learning of Sounding Object Visual Grounding and Sound Separation
Yapeng Tian, Di Hu, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2021, pp. 2745-2754
There are rich synchronized audio and visual events in our daily life. Inside th e events, audio scenes are associated with the corresponding visual objects; mea nwhile, sounding objects can indicate and help to separate their individual soun

ds in the audio track. Based on this observation, in this paper, we propose a cy
clic co-learning (CCoL) paradigm that can jointly learn sounding object visual g
rounding and audio-visual sound separation in a unified framework. Concretely, w
e can leverage grounded object-sound relations to improve the results of sound s
eparation. Meanwhile, benefiting from discriminative information from separated
sounds, we improve training example sampling for sounding object grounding, whic
h builds a co-learning cycle for the two tasks and makes them mutually beneficia
l. Extensive experiments show that the proposed framework outperforms the compar
ed recent approaches on both tasks, and they can benefit from each other with ou
r cyclic co-learning. The source code and pre-trained models are released in htt
ps://github.com/YapengTian/CCOL-CVPR21.
****
Digital Gimbal: End-to-End Deep Image Stabilization With Learnable Exposure Time
s
Omer Dahary, Matan Jacoby, Alex M. Bronstein; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11936-11945
Mechanical image stabilization using actuated gimbals enables capturing long-exp
osure shots without suffering from blur due to camera motion. These devices, how
ever, are often physically cumbersome and expensive, limiting their widespread u
se. In this work, we propose to digitally emulate a mechanically stabilized syst
em from the input of a fast unstabilized camera. To exploit the trade-off betwee
n motion blur at long exposures and low SNR at short exposures, we train a CNN t
hat estimates a sharp high-SNR image by aggregating a burst of noisy short-expos
ure frames, related by unknown motion. We further suggest learning the burst's e
xposure times in an end-to-end manner, thus balancing the noise and blur across
the frames. We demonstrate this method's advantage over the traditional approach
 of deblurring a single image or denoising a fixed-exposure burst on both synthe
tic and real data.
****
Rethinking Text Segmentation: A Novel Dataset and a Text-Specific Refinement App
roach
Xingqian Xu, Zhifei Zhang, Zhaowen Wang, Brian Price, Zhonghao Wang, Humphrey Sh
i; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2021, pp. 12045-12055
Text segmentation is a prerequisite in many real-world text-related tasks, e.g.,
 text style transfer, and scene text removal. However, facing the lack of high-q
uality datasets and dedicated investigations, this critical prerequisite has bee
n left as an assumption in many works, and has been largely overlooked by curren
t research. To bridge this gap, we proposed TextSeg, a large-scale fine-annotate
d text dataset with six types of annotations: word- and character-wise bounding
polygons, masks, and transcriptions. We also introduce Text Refinement Network (
TexRNet), a novel text segmentation approach that adapts to the unique propertie
s of text, e.g. non-convex boundary, diverse texture, etc., which often impose b
urdens on traditional segmentation models. In our TexRNet, we propose text-speci
fic network designs to address such challenges, including key features pooling a
nd attention-based similarity checking. We also introduce trimap and discriminat
or losses that show significant improvement in text segmentation. Extensive expe
riments are carried out on both our TextSeg dataset and other existing datasets.
 We demonstrate that TexRNet consistently improves text segmentation performance
 by nearly 2% compared to other state-of-the-art segmentation methods. Our datas
et and code can be found at https://github.com/SHI-Labs/Rethinking-Text-Segmenta
tion.
****
SUTD-TrafficQA: A Question Answering Benchmark and an Efficient Network for Vide
o Reasoning Over Traffic Events
Li Xu, He Huang, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2021, pp. 9878-9888
Traffic event cognition and reasoning in videos is an important task that has a
wide range of applications in intelligent transportation, assisted driving, and
autonomous vehicles. In this paper, we create a novel dataset, SUTD-TrafficQA (T

raffic Question Answering), which takes the form of video QA based on the collected 10,080 in-the-wild videos and annotated 62,535 QA pairs, for benchmarking the cognitive capability of causal inference and event understanding models in complex traffic scenarios. Specifically, we propose 6 challenging reasoning tasks corresponding to various traffic scenarios, so as to evaluate the reasoning capability over different kinds of complex yet practical traffic events. Moreover, we propose Eclipse, a novel Efficient glimpse network via dynamic inference, in order to achieve computation-efficient and reliable video reasoning. The experiments show that our method achieves superior performance while reducing the computation cost significantly.
****

T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval
Xiaohan Wang, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5079-5088
Text-video retrieval is a challenging task that aims to search relevant video contents based on natural language descriptions. The key to this problem is to measure text-video similarities in a joint embedding space. However, most existing methods only consider the global cross-modal similarity and overlook the local details. Some works incorporate the local comparisons through cross-modal local matching and reasoning. These complex operations introduce tremendous computation. In this paper, we design an efficient global-local alignment method. The multi-modal video sequences and text features are adaptively aggregated with a set of shared semantic centers. The local cross-modal similarities are computed between the video feature and text feature within the same center. This design enables the meticulous local comparison and reduces the computational cost of the interaction between each text-video pair. Moreover, a global alignment method is proposed to provide a global cross-modal measurement that is complementary to the local perspective. The global aggregated visual features also provide additional supervision, which is indispensable to the optimization of the learnable semantic centers. We achieve consistent improvements on three standard text-video retrieval benchmarks and outperform the state-of-the-art by a clear margin.
****

Privacy-Preserving Image Features via Adversarial Affine Subspace Embeddings
Mihai Dusmanu, Johannes L. Schonberger, Sudipta N. Sinha, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14267-14277
Many computer vision systems require users to upload image features to the cloud for processing and storage. These features can be exploited to recover sensitive information about the scene or subjects, e.g., by reconstructing the appearance of the original image. To address this privacy concern, we propose a new privacy-preserving feature representation. The core idea of our work is to drop constraints from each feature descriptor by embedding it within an affine subspace containing the original feature as well as adversarial feature samples. Feature matching on the privacy-preserving representation is enabled based on the notion of subspace-to-subspace distance. We experimentally demonstrate the effectiveness of our method and its high practical relevance for the applications of visual localization and mapping as well as face authentication. Compared to the original features, our approach makes it significantly more difficult for an adversary to recover private information.
****

StyleMeUp: Towards Style-Agnostic Sketch-Based Image Retrieval
Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8504-8513
Sketch-based image retrieval (SBIR) is a cross-modal matching problem which is typically solved by learning a joint embedding space where the semantic content shared between photo and sketch modalities are preserved. However, a fundamental challenge in SBIR has been largely ignored so far, that is, sketches are drawn by humans and considerable style variations exist amongst different users. An effective SBIR model needs to explicitly account for this style diversity, cruciall

y, to generalise to unseen user styles. To this end, a novel style-agnostic SBIR model is proposed. Different from existing models, a cross-modal variational autoencoder (VAE) is employed to explicitly disentangle each sketch into a semantic content part shared with the corresponding photo, and a style part unique to the sketcher. Importantly, to make our model dynamically adaptable to any unseen user styles, we propose to meta-train our cross-modal VAE by adding two style-adaptive components: a set of feature transformation layers to its encoder and a regulariser to the disentangled semantic content latent code. With this meta-learning framework, our model can not only disentangle the cross-modal shared semantic content for SBIR, but can adapt the disentanglement to any unseen user style as well, making the SBIR model truly style-agnostic. Extensive experiments show that our style-agnostic model yields state-of-the-art performance for both category-level and instance-level SBIR.
****
Embedding Transfer With Label Relaxation for Improved Metric Learning
Sungyeon Kim, Dongwon Kim, Minsu Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3967-3976
This paper presents a novel method for embedding transfer, a task of transferring knowledge of a learned embedding model to another. Our method exploits pairwise similarities between samples in the source embedding space as the knowledge, and transfers them through a loss used for learning target embedding models. To this end, we design a new loss called relaxed contrastive loss, which employs the pairwise similarities as relaxed labels for inter-sample relations. Our loss provides a rich supervisory signal beyond class equivalence, enables more important pairs to contribute more to training, and imposes no restriction on manifolds of target embedding spaces. Experiments on metric learning benchmarks demonstrate that our method largely improves performance, or reduces sizes and output dimensions of target models effectively. We further show that it can be also used to enhance quality of self-supervised representation and performance of classification models. In all the experiments, our method clearly outperforms existing embedding transfer techniques.
****
Beyond Static Features for Temporally Consistent 3D Human Pose and Shape From a Video
Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1964-1973
Despite the recent success of single image-based 3D human pose and shape estimation methods, recovering temporally consistent and smooth 3D human motion from a video is still challenging. Several video-based methods have been proposed; however, they fail to resolve the single image-based methods' temporal inconsistency issue due to a strong dependency on a static feature of the current frame. In this regard, we present a temporally consistent mesh recovery system (TCMR). It effectively focuses on the past and future frames' temporal information without being dominated by the current static feature. Our TCMR significantly outperforms previous video-based methods in temporal consistency with better per-frame 3D pose and shape accuracy. We also release the codes.
****
Layout-Guided Novel View Synthesis From a Single Indoor Panorama
Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16438-16447
Existing view synthesis methods mainly focus on the perspective images and have shown promising results. However, due to the limited field-of-view of the pinhole camera, the performance quickly degrades when large camera movements are adopted. In this paper, we make the first attempt to generate novel views from a single indoor panorama and take the large camera translations into consideration. To tackle this challenging problem, we first use Convolutional Neural Networks (CNNs) to extract the deep features and estimate the depth map from the source-view image. Then, we leverage the room layout prior, a strong structural constraint

of the indoor scene, to guide the generation of target views. More concretely, we estimate the room layout in the source view and transform it into the target viewpoint as guidance. Meanwhile, we also constrain the room layout of the generated target-view images to enforce geometric consistency. To validate the effectiveness of our method, we further build a large-scale photo-realistic dataset containing both small and large camera translations. The experimental results on our challenging dataset demonstrate that our method achieves state-of-the-art performance. The project page is at https://github.com/bluestyle97/PNVS.
****
STMTrack: Template-Free Visual Tracking With Space-Time Memory Networks
Zhihong Fu, Qingjie Liu, Zehua Fu, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13774-13783
Boosting performance of the offline trained siamese trackers is getting harder nowadays since the fixed information of the template cropped from the first frame has been almost thoroughly mined, but they are poorly capable of resisting target appearance changes. Existing trackers with template updating mechanisms rely on time-consuming numerical optimization and complex hand-designed strategies to achieve competitive performance, hindering them from real-time tracking and practical applications. In this paper, we propose a novel tracking framework built on top of a space-time memory network that is competent to make full use of historical information related to the target for better adapting to appearance variations during tracking. Specifically, a novel memory mechanism is introduced, which stores the historical information of the target to guide the tracker to focus on the most informative regions in the current frame. Furthermore, the pixel-level similarity computation of the memory network enables our tracker to generate much more accurate bounding boxes of the target. Extensive experiments and comparisons with many competitive trackers on challenging large-scale benchmarks, OTB-2015, TrackingNet, GOT-10k, LaSOT, UAV123, and VOT2018, show that, without bells and whistles, our tracker outperforms all previous state-of-the-art real-time methods while running at 37 FPS. The code is available at https://github.com/fzh0917/STMTrack.
****
Reformulating HOI Detection As Adaptive Set Prediction
Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, Chen Qian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9004-9013
Determining which image regions to concentrate is critical for Human-Object Interaction (HOI) detection. Conventional HOI detectors focus on either detected human and object pairs or pre-defined interaction locations, which limits learning of the effective features. In this paper, we reformulate HOI detection as an adaptive set prediction problem, with this novel formulation, we propose an Adaptive Set-based one-stage framework (AS-Net) with parallel instance and interaction branches. To attain this, we map a trainable interaction query set to an interaction prediction set with transformer. Each query adaptively aggregates the interaction-relevant features from global contexts through multi-head co-attention. Besides, the training process is supervised adaptively by matching each ground-truth with the interaction prediction. Furthermore, we design an effective instance-aware attention module to introduce instructive features from the instance branch into the interaction branch. Our method outperforms previous state-of-the-art methods without any extra human pose and language features on three challenging HOI detection datasets. Especially, we achieve over 31% relative improvement on a large scale HICO-DET dataset. Code is available at https://github.com/yoyomimi/AS-Net.
****
Strengthen Learning Tolerance for Weakly Supervised Object Localization
Guangyu Guo, Junwei Han, Fang Wan, Dingwen Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7403-7412
Weakly supervised object localization (WSOL) aims at learning to localize objects of interest by only using the image-level labels as the supervision. While numerous efforts have been made in this field, recent approaches still suffer from

two challenges: one is the part domination issue while the other is the learning robustness issue. Specifically, the former makes the localizer prone to the local discriminative object regions rather than the desired whole object, and the latter makes the localizer over-sensitive to the variations of the input images so that one can hardly obtain localization results robust to the arbitrary visual stimulus. To solve these issues, we propose a novel framework to strengthen the learning tolerance, referred to as SLT-Net, for WSOL. Specifically, we consider two-fold learning tolerance strengthening mechanisms. One is the semantic tolerance strengthening mechanism, which allows the localizer to make mistakes for classifying similar semantics so that it will not concentrate too much on the discriminative local regions. The other is the visual stimuli tolerance strengthening mechanism, which enforces the localizer to be robust to different image transformations so that the prediction quality will not be sensitive to each specific input image. Finally, we implement comprehensive experimental comparisons on two widely-used datasets CUB and ILSVRC2012, which demonstrate the effectiveness of our proposed approach.
****
Mesh Saliency: An Independent Perceptual Measure or a Derivative of Image Saliency?
Ran Song, Wei Zhang, Yitian Zhao, Yonghuai Liu, Paul L. Rosin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8853-8862
While mesh saliency aims to predict regional importance of 3D surfaces in agreement with human visual perception and is well researched in computer vision and graphics, latest work with eye-tracking experiments shows that state-of-the-art mesh saliency methods remain poor at predicting human fixations. Cues emerging prominently from these experiments suggest that mesh saliency might associate with the saliency of 2D natural images. This paper proposes a novel deep neural network for learning mesh saliency using image saliency ground truth to 1) investigate whether mesh saliency is an independent perceptual measure or just a derivative of image saliency and 2) provide a weakly supervised method for more accurately predicting mesh saliency. Through extensive experiments, we not only demonstrate that our method outperforms the current state-of-the-art mesh saliency method by 116% and 21% in terms of linear correlation coefficient and AUC respectively, but also reveal that mesh saliency is intrinsically related with both image saliency and object categorical information. Codes are available at https://github.com/rsong/MIMO-GAN.
****
Passive Inter-Photon Imaging
Atul Ingle, Trevor Seets, Mauro Buttafava, Shantanu Gupta, Alberto Tosi, Mohit Gupta, Andreas Velten; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8585-8595
Digital camera pixels measure image intensities by converting incident light energy into an analog electrical current, and then digitizing it into a fixed-width binary representation. This direct measurement method, while conceptually simple, suffers from limited dynamic range and poor performance under extreme illumination --- electronic noise dominates under low illumination, and pixel full-well capacity results in saturation under bright illumination. We propose a novel intensity cue based on measuring inter-photon timing, defined as the time delay between detection of successive photons. Based on the statistics of inter-photon times measured by a time-resolved single-photon sensor, we develop theory and algorithms for a scene brightness estimator which works over extreme dynamic range; we experimentally demonstrate imaging scenes with a dynamic range of over ten million to one. The proposed techniques, aided by the emergence of single-photon sensors such as single-photon avalanche diodes (SPADs) with picosecond timing resolution, will have implications for a wide range of imaging applications: robotics, consumer photography, astronomy, microscopy and biomedical imaging.
****
Domain Consensus Clustering for Universal Domain Adaptation
Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, Yi Yang; Proceedings of the IEE

E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9757-9766

In this paper, we investigate Universal Domain Adaptation (UniDA) problem, which aims to transfer the knowledge from source to target under unaligned label space. The main challenge of UniDA lies in how to separate common classes (i.e., classes shared across domains), from private classes (i.e., classes only exist in one domain). Previous works treat the private samples in the target as one generic class but ignore their intrinsic structure. Consequently, the resulting representations are not compact enough in the latent space and can be easily confused with common samples. To better exploit the intrinsic structure of the target domain, we propose Domain Consensus Clustering(DCC), which exploits the domain consensus knowledge to discover discriminative clusters on both common samples and private ones. Specifically, we draw the domain consensus knowledge from two aspects to facilitate the clustering and the private class discovery, i.e., the semantic-level consensus, which identifies the cycle-consistent clusters as the common classes, and the sample-level consensus, which utilizes the cross-domain classification agreement to determine the number of clusters and discover the private classes. Based on DCC, we are able to separate the private classes from the common ones, and differentiate the private classes themselves. Finally, we apply a class-aware alignment technique on identified common samples to minimize the distribution shift, and a prototypical regularizer to inspire discriminative target clusters. Experiments on four benchmarks demonstrate DCC significantly outperforms previous state-of-the-arts.

****

Continual Semantic Segmentation via Repulsion-Attraction of Sparse and Disentangled Latent Representations

Umberto Michieli, Pietro Zanuttigh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1114-1124

Deep neural networks suffer from the major limitation of catastrophic forgetting old tasks when learning new ones. In this paper we focus on class incremental continual learning in semantic segmentation, where new categories are made available over time while previous training data is not retained. The proposed continual learning scheme shapes the latent space to reduce forgetting whilst improving the recognition of novel classes. Our framework is driven by three novel components which we also combine on top of existing techniques effortlessly. First, prototypes matching enforces latent space consistency on old classes, constraining the encoder to produce similar latent representation for previously seen classes in the subsequent steps. Second, features sparsification allows to make room in the latent space to accommodate novel classes. Finally, contrastive learning is employed to cluster features according to their semantics while tearing apart those of different classes. Extensive evaluation on the Pascal VOC2012 and ADE20K datasets demonstrates the effectiveness of our approach, significantly outperforming state-of-the-art methods.

****

Audio-Driven Emotional Video Portraits

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14080-14089

Despite previous success in generating audio-driven talking heads, most of the previous studies focus on the correlation between speech content and the mouth shape. Facial emotion, which is one of the most important features on natural human faces, is always neglected in their methods. In this work, we present Emotional Video Portraits (EVP), a system for synthesizing high-quality video portraits with vivid emotional dynamics driven by audios. Specifically, we propose the Cross-Reconstructed Emotion Disentanglement technique to decompose speech into two decoupled spaces, i.e., a duration-independent emotion space and a duration dependent content space. With the disentangled features, dynamic 2D emotional facial landmarks can be deduced. Then we propose the Target-Adaptive Face Synthesis technique to generate the final high-quality video portraits, by bridging the gap between the deduced landmarks and the natural head poses of target videos. Exten

sive experiments demonstrate the effectiveness of our method both qualitatively and quantitatively.
****
Pareto Self-Supervised Training for Few-Shot Learning
Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, Donglin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13663-13672
While few-shot learning (FSL) aims for rapid generalization to new concepts with little supervision, self-supervised learning (SSL) constructs supervisory signals directly computed from unlabeled data. Exploiting the complementarity of these two manners, few-shot auxiliary learning has recently drawn much attention to deal with few labeled data. Previous works benefit from sharing inductive bias between the main task (FSL) and auxiliary tasks (SSL), where the shared parameters of tasks are optimized by minimizing a linear combination of task losses. However, it is challenging to select a proper weight to balance tasks and reduce task conflict. To handle the problem as a whole, we propose a novel approach named as Pareto self-supervised training (PSST) for FSL. PSST explicitly decomposes the few-shot auxiliary problem into multiple constrained multi-objective subproblems with different trade-off preferences, and here a preference region in which the main task achieves the best performance is identified. Then, an effective preferred Pareto exploration is proposed to find a set of optimal solutions in such a preference region. Extensive experiments on several public benchmark datasets validate the effectiveness of our approach by achieving state-of-the-art performance.
****
EnD: Entangling and Disentangling Deep Representations for Bias Correction
Enzo Tartaglione, Carlo Alberto Barbano, Marco Grangetto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13508-13517
Artificial neural networks perform state-of-the-art in an ever-growing number of tasks, and nowadays they are used to solve an incredibly large variety of tasks. There are problems, like the presence of biases in the training data, which question the generalization capability of these models. In this work we propose EnD, a regularization strategy whose aim is to prevent deep models from learning unwanted biases. In particular, we insert an ""information bottleneck"" at a certain point of the deep neural network, where we disentangle the information about the bias, still letting the useful information for the training task forward-propagating in the rest of the model. One big advantage of EnD is that it does not require additional training complexity (like decoders or extra layers in the model), since it is a regularizer directly applied on the trained model. Our experiments show that EnD effectively improves the generalization on unbiased test sets, and it can be effectively applied on real-case scenarios, like removing hidden biases in the COVID-19 detection from radiographic images.
****
Recorrupted-to-Recorrupted: Unsupervised Deep Learning for Image Denoising
Tongyao Pang, Huan Zheng, Yuhui Quan, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2043-2052
Deep denoiser, the deep network for denoising, has been the focus of the recent development on image denoising. In the last few years, there is an increasing interest in developing unsupervised deep denoisers which only call unorganized noisy images without ground truth for training. Nevertheless, the performance of these unsupervised deep denoisers is not competitive to their supervised counterparts. Aiming at developing a more powerful unsupervised deep denoiser, this paper proposed a data augmentation technique, called recorrupted-to-recorrupted (R2R), to address the overfitting caused by the absence of truth images. For each noisy image, we showed that the cost function defined on the noisy/noisy image pairs constructed by the R2R method is statistically equivalent to its supervised counterpart defined on the noisy/truth image pairs. Extensive experiments showed that the proposed R2R method noticeably outperformed existing unsupervised deep denoisers, and is competitive to representative supervised deep denoisers.

****
Reconsidering Representation Alignment for Multi-View Clustering
Daniel J. Trosten, Sigurd Lokse, Robert Jenssen, Michael Kampffmeyer; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1255-1265
Aligning distributions of view representations is a core component of today's st ate of the art models for deep multi-view clustering. However, we identify sever al drawbacks with naively aligning representation distributions. We demonstrate that these drawbacks both lead to less separable clusters in the representation space, and inhibit the model's ability to prioritize views. Based on these obser vations, we develop a simple baseline model for deep multi-view clustering. Our baseline model avoids representation alignment altogether, while performing simi lar to, or better than, the current state of the art. We also expand our baselin e model by adding a contrastive learning component. This introduces a selective alignment procedure that preserves the model's ability to prioritize views. Our experiments show that the contrastive learning component enhances the baseline m odel, improving on the current state of the art by a large margin on several dat asets.
****
Probabilistic Embeddings for Cross-Modal Retrieval
Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, Dian e Larlus; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8415-8424
Cross-modal retrieval methods build a common representation space for samples fr om multiple modalities, typically from the vision and the language domains. For images and their captions, the multiplicity of the correspondences makes the tas k particularly challenging. Given an image (respectively a caption), there are m ultiple captions (respectively images) that equally make sense. In this paper, w e argue that deterministic functions are not sufficiently powerful to capture su ch one-to-many correspondences. Instead, we propose to use Probabilistic Cross-M odal Embedding (PCME), where samples from the different modalities are represent ed as probabilistic distributions in the common embedding space. Since common be nchmarks such as COCO suffer from non-exhaustive annotations for cross-modal mat ches, we propose to additionally evaluate retrieval on the CUB dataset, a smalle r yet clean database where all possible image-caption pairs are annotated. We ex tensively ablate PCME and demonstrate that it not only improves the retrieval pe rformance over its deterministic counterpart but also provides uncertainty estim ates that render the embeddings more interpretable. Code is available at https:/ /github.com/naver-ai/pcme.
****
Cloud2Curve: Generation and Vectorization of Parametric Sketches
Ayan Das, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yi-Zhe Song; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7088-7097
Analysis of human sketches in deep learning has advanced immensely through the u se of waypoint-sequences rather than raster-graphic representations. We further aim to model sketches as a sequence of low-dimensional parametric curves. To thi s end, we propose an inverse graphics framework capable of approximating a raste r or waypoint based stroke encoded as a point-cloud with a variable-degree Bezie r curve. Building on this module, we present Cloud2Curve, a generative model for scalable high-resolution vector sketches that can be trained end-to-end using p oint-cloud data alone. As a consequence, our model is also capable of determinis tic vectorization which can map novel raster or waypoint based sketches to their corresponding high-resolution scalable Bezier equivalent. We evaluate the gener ation and vectorization capabilities of our model on Quick, Draw! and K-MNIST da tasets.
****
TransFill: Reference-Guided Image Inpainting by Merging Multiple Color and Spati al Transformations
Yuqian Zhou, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi; Proceedings of t

he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2266-2276
Image inpainting is the task of plausibly restoring missing pixels within a hole region that is to be removed from a target image. Most existing technologies exploit patch similarities within the image, or leverage large-scale training data to fill the hole using learned semantic and texture information. However, due to the ill-posed nature of the inpainting task, such methods struggle to complete larger holes containing complicated scenes. In this paper, we propose TransFill, a multi-homography transformed fusion method to fill the hole by referring to another source image that shares scene contents with the target image. We first align the source image to the target image by estimating multiple homographies guided by different depth levels. We then learn to adjust the color and apply a pixel-level warping to each homography-warped source image to make it more consistent with the target. Finally, a pixel-level fusion module is learned to selectively merge the different proposals. Our method achieves state-of-the-art performance on pairs of images across a variety of wide baselines and color differences, and generalizes to user-provided image pairs.
****
On Focal Loss for Class-Posterior Probability Estimation: A Theoretical Perspective
Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, Masashi Sugiyama; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5202-5211
The focal loss has demonstrated its effectiveness in many real-world applications such as object detection and image classification, but its theoretical understanding has been limited so far. In this paper, we first prove that the focal loss is classification-calibrated, i.e., its minimizer surely yields the Bayes-optimal classifier and thus the use of the focal loss in classification can be theoretically justified. However, we also prove a negative fact that the focal loss is not strictly proper, i.e., the confidence score of the classifier obtained by focal loss minimization does not match the true class-posterior probability. This may cause the trained classifier to give an unreliable confidence score, which can be harmful in critical applications. To mitigate this problem, we prove that there exists a particular closed-form transformation that can recover the true class-posterior probability from the outputs of the focal risk minimizer. Our experiments show that our proposed transformation successfully improves the quality of class-posterior probability estimation and improves the calibration of the trained classifier, while preserving the same prediction accuracy.
****
VIP-DeepLab: Learning Visual Perception With Depth-Aware Video Panoptic Segmentation
Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3997-4008
In this paper, we present ViP-DeepLab, a unified model attempting to tackle the long-standing and challenging inverse projection problem in vision, which we model as restoring the point clouds from perspective image sequences while providing each point with instance-level semantic interpretations. Solving this problem requires the vision models to predict the spatial location, semantic class, and temporally consistent instance label for each 3D point. ViP-DeepLab approaches it by jointly performing monocular depth estimation and video panoptic segmentation. We name this joint task as Depth-aware Video Panoptic Segmentation, and propose a new evaluation metric along with two derived datasets for it, which will be made available to the public. On the individual sub-tasks, ViP-DeepLab also achieves state-of-the-art results, outperforming previous methods by 5.1% VPQ on Cityscapes-VPS, ranking 1st on the KITTI monocular depth estimation benchmark, and 1st on KITTI MOTS pedestrian. The datasets and the evaluation codes are made publicly available.
****
Sequence-to-Sequence Contrastive Learning for Text Recognition

Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R. Manmatha, Pietro Perona; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15302-15312
We propose a framework for sequence-to-sequence contrastive learning (SeqCLR) of visual representations, which we apply to text recognition. To account for the sequence-to-sequence structure, each feature map is divided into different instances over which the contrastive loss is computed. This operation enables us to contrast in a sub-word level, where from each image we extract several positive pairs and multiple negative examples. To yield effective visual representations for text recognition, we further suggest novel augmentation heuristics, different encoder architectures and custom projection heads. Experiments on handwritten text and on scene text show that when a text decoder is trained on the learned representations, our method outperforms non-sequential contrastive methods. In addition, when the amount of supervision is reduced, SeqCLR significantly improves performance compared with supervised training, and when fine-tuned with 100% of the labels, our method achieves state-of-the-art results on standard handwritten text recognition benchmarks.
****
Prototype-Supervised Adversarial Network for Targeted Attack of Deep Hashing
Xunguang Wang, Zheng Zhang, Baoyuan Wu, Fumin Shen, Guangming Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16357-16366
Due to its powerful capability of representation learning and high-efficiency computation, deep hashing has made significant progress in large-scale image retrieval. However, deep hashing networks are vulnerable to adversarial examples, which is a practical secure problem but seldom studied in hashing-based retrieval field. In this paper, we propose a novel prototype-supervised adversarial network (ProS-GAN), which formulates a flexible generative architecture for efficient and effective targeted hashing attack. To the best of our knowledge, this is the first generation-based method to attack deep hashing networks. Generally, our proposed framework consists of three parts, i.e., a PrototypeNet, a generator and a discriminator. Specifically, the designed PrototypeNet embeds the target label into the semantic representation and learns the prototype code as the category-level representative of the target label. Moreover, the semantic representation and the original image are jointly fed into the generator for flexible targeted attack. Particularly, the prototype code is adopted to supervise the generator to construct the targeted adversarial example by minimizing the Hamming distance between the hash code of the adversarial example and the prototype code. Furthermore, the generator is against the discriminator to simultaneously encourage the adversarial examples visually realistic and the semantic representation informative. Extensive experiments verify that the proposed framework can efficiently produce adversarial examples with better targeted attack performance and transferability over state-of-the-art targeted attack methods of deep hashing.
****
PD-GAN: Probabilistic Diverse GAN for Image Inpainting
Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, Jing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9371-9381
We propose PD-GAN, a probabilistic diverse GAN forimage inpainting. Given an input image with arbitrary holeregions, PD-GAN produces multiple inpainting results withdiverse and visually realistic content. Our PD-GAN is builtupon a vanilla GAN which generates images based on random noise. During image generation, we modulate deepfeatures of input random noise from coarse-to-fine by injecting an initially restored image and the hole regions inmultiple scales. We argue that during hole filling, the pixels near the hole boundary should be more deterministic(i.e., with higher probability trusting the context and initially restored image to create natural inpainting boundary), while those pixels lie in the center of the hole shouldenjoy more degrees of freedom (i.e., more likely to dependon the random noise for enhancing diversity). To this end, we propose spatially probabilistic diversity normalization(SPDNorm) inside the modulation to model the proba

bilityof generating a pixel conditioned on the context information. SPDNorm dyna mically balances the realism and diversity inside the hole region, making the ge nerated content more diverse towards the hole center and resembleneighboring ima ge content more towards the hole boundary. Meanwhile, we propose a perceptual di versity loss tofurther empower PD-GAN for diverse content generation. Experiment s on benchmark datasets including CelebA-HQ, Places2 and Paris Street View indic ate that PD-GAN is ef-fective for diverse and visually realistic image restorati on.
****
Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation
Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, Barret Zoph; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2021, pp. 2918-2928
Building instance segmentation models that are data-efficient and can handle rar e object categories is an important challenge in computer vision. Leveraging dat a augmentations is a promising direction towards addressing this challenge. Here , we perform a systematic study of the Copy-Paste augmentation (e.g., [13, 12]) for instance segmentation where we randomly paste objects onto an image. Prior s tudies on Copy-Paste relied on modeling the surrounding visual context for pasti ng the objects. However, we find that the simple mechanism of pasting objects ra ndomly is good enough and can provide solid gains on top of strong baselines. Fu rthermore, we show Copy-Paste is additive with semi-supervised methods that leve rage extra data through pseudo labeling (eg. self-training). On COCO instance se gmentation, we achieve 49.1 mask AP and 57.3 box AP, an improvement of +0.6 mask AP and +1.5 box AP over the previous state-of-the-art. We further demonstrate t hat Copy-Paste can lead to significant improvements on the LVIS benchmark. Our b aseline model outperforms the LVIS 2020 Challenge winning entry by +3.6 mask AP on rare categories.
****
Learning Deep Latent Variable Models by Short-Run MCMC Inference With Optimal Tr ansport Correction
Dongsheng An, Jianwen Xie, Ping Li; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2021, pp. 15415-15424
Learning latent variable models with deep top-down architectures typically requi res inferring the latent variables for each training example based on the poster ior distribution of these latent variables. The inference step typically relies on either time-consuming long run Markov chain Monte Caro (MCMC) or a separate i nference model for variational learning. In this paper, we propose to use short run MCMC, such as Langevin dynamics, as an approximate inference engine, where t he bias existing in the output distribution of the short run Langevin dynamics i s corrected by optimal transport, which aims at minimizing the Wasserstein dista nce between the biased distribution produced by the finite step Langevin dynamic s and the prior distribution. Our experiments show that the proposed strategy ou tperforms the variational auto-encoder (VAE) and alternating back-propagation al gorithm (ABP) in terms of reconstruction error and synthesis quality.
****
MobileDets: Searching for Object Detection Architectures for Mobile Accelerators
Yunyang Xiong, Hanxiao Liu, Suyog Gupta, Berkin Akin, Gabriel Bender, Yongzhe Wa ng, Pieter-Jan Kindermans, Mingxing Tan, Vikas Singh, Bo Chen; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 3825-3834
Inverted bottleneck layers, which are built upon depthwise convolutions, have be en the predominant building blocks in state-of-the-art object detection models o n mobile devices. In this work, we investigate the optimality of this design pat tern over a broad range of mobile accelerators by revisiting the usefulness of r egular convolutions. We discover that regular convolutions are a potent componen t to boost the latency-accuracy trade-off for object detection on accelerators, provided that they are placed strategically in the network via neural architectu re search. By incorporating regular convolutions in the search space and directl y optimizing the network architectures for object detection, we obtain a family

of object detection models, MobileDets, that achieve state-of-the-art results ac ross mobile accelerators. On the COCO object detection task, MobileDets outperfo rm MobileNetV3+SSDLite by 1.7 mAP at comparable mobile CPU inference latencies. MobileDets also outperform MobileNetV2+SSDLite by 1.9 mAP on mobile CPUs, 3.7 mA P on Google EdgeTPU, 3.4 mAP on Qualcomm Hexagon DSP and 2.7 mAP on Nvidia Jetso n GPU without increasing latency. Moreover, MobileDets are comparable with the s tate-of-the-art MnasFPN on mobile CPUs even without using the feature pyramid, a nd achieve better mAP scores on both EdgeTPUs and DSPs with up to 2x speedup. Co de and models are available in the TensorFlow Object Detection API: https://gith ub.com/tensorflow/models/tree/master/research/object_detection.
****
Self-Supervised Geometric Perception
Heng Yang, Wei Dong, Luca Carlone, Vladlen Koltun; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14350-143 61
We present self-supervised geometric perception (SGP), the first general framewo rk to learn a feature descriptor for correspondence matching without any ground-truth geometric model labels (e.g., camera poses, rigid transformations). Our fi rst contribution is to formulate geometric perception as an optimization problem that jointly optimizes the feature descriptor and the geometric models given a large corpus of visual measurements (e.g., images, point clouds). Under this opt imization formulation, we show that two important streams of research in vision, namely robust model fitting and deep feature learning, correspond to optimizing one block of the unknown variables while fixing the other block. This analysis naturally leads to our second contribution - the SGP algorithm that performs alt ernating minimization to solve the joint optimization. SGP iteratively executes two meta-algorithms: a teacher that performs robust model fitting given learned features to generate geometric pseudo-labels, and a student that performs deep f eature learning under noisy supervision of the pseudo-labels. As a third contrib ution, we apply SGP to two perception problems on large-scale real datasets, nam ely relative camera pose estimation on MegaDepth and point cloud registration on 3DMatch. We demonstrate that SGP achieves state-of-the-art performance that is on-par or superior to the supervised oracles trained using ground-truth labels.
****
CutPaste: Self-Supervised Learning for Anomaly Detection and Localization
Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, Tomas Pfister; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 966 4-9674
We aim at constructing a high performance model for defect detection that detect s unknown anomalous patterns of an image without anomalous data. To this end, we propose a two-stage framework for building anomaly detectors using normal train ing data only. We first learn self-supervised deep representations and then buil d a generative one-class classifier on learned representations. We learn represe ntations by classifying normal data from the CutPaste, a simple data augmentatio n strategy that cuts an image patch and pastes at a random location of a large i mage. Our empirical study on MVTec anomaly detection dataset demonstrates the pr oposed algorithm is general to be able to detect various types of real-world def ects. We bring the improvement upon previous arts by 3.1 AUCs when learning repr esentations from scratch. By transfer learning on pretrained representations on ImageNet, we achieve a new state-of-the-art 96.6 AUC. Lastly, we extend the fram ework to learn and extract representations from patches to allow localizing defe ctive areas without annotations during training.
****
Open World Compositional Zero-Shot Learning
Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, Zeynep Akata; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5222-5230
Compositional Zero-Shot learning (CZSL) requires to recognize state-object compo sitions unseen during training. In this work, instead of assuming prior knowledg e about the unseen compositions, we operate in the open world setting, where the

search space includes a large number of unseen compositions some of which might be unfeasible. In this setting, we start from the cosine similarity between visual features and compositional embeddings. After estimating the feasibility score of each composition, we use these scores to either directly mask the output space or as a margin for the cosine similarity between visual features and compositional embeddings during training. Our experiments on two standard CZSL benchmarks show that all the methods suffer severe performance degradation when applied in the open world setting. While our simple CZSL model achieves state-of-the-art performances in the closed world scenario, our feasibility scores boost the performance of our approach in the open world setting, clearly outperforming the previous state of the art. Code is available at: https://github.com/ExplainableML/czsl.
****
Bi-GCN: Binary Graph Convolutional Network
Junfu Wang, Yunhong Wang, Zhen Yang, Liang Yang, Yuanfang Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1561-1570
Graph Neural Networks (GNNs) have achieved tremendous success in graph representation learning. Unfortunately, current GNNs usually rely on loading the entire attributed graph into network for processing. This implicit assumption may not be satisfied with limited memory resources, especially when the attributed graph is large. In this paper, we pioneer to propose a Binary Graph Convolutional Network (Bi-GCN), which binarizes both the network parameters and input node features. Besides, the original matrix multiplications are revised to binary operations for accelerations. According to the theoretical analysis, our Bi-GCN can reduce the memory consumption by an average of 30x for both the network parameters and input data, and accelerate the inference speed by an average of 47x, on the citation networks. Meanwhile, we also design a new gradient approximation based back-propagation method to train our Bi-GCN well. Extensive experiments have demonstrated that our Bi-GCN can give a comparable performance compared to the full-precision baselines. Besides, our binarization approach can be easily applied to other GNNs, which has been verified in the experiments.
****
Complementary Relation Contrastive Distillation
Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, Xiaohua Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9260-9269
Knowledge distillation aims to transfer representation ability from a teacher model to a student model. Previous approaches focus on either individual representation distillation or inter-sample similarity preservation. While we argue that the inter-sample relation conveys abundant information and needs to be distilled in a more effective way. In this paper, we propose a novel knowledge distillation method, namely Complementary Relation Contrastive Distillation (CRCD), to transfer the structural knowledge from the teacher to the student. Specifically, we estimate the mutual relation in an anchor-based way and distill the anchor-student relation under the supervision of its corresponding anchor-teacher relation. To make it more robust, mutual relations are modeled by two complementary elements: the feature and its gradient. Furthermore, the low bound of mutual information between the anchor-teacher relation distribution and the anchor-student relation distribution is maximized via relation contrastive loss, which can distill both the sample representation and the inter-sample relations. Experiments on different benchmarks demonstrate the effectiveness of our proposed CRCD.
****
UnrealPerson: An Adaptive Pipeline Towards Costless Person Re-Identification
Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11506-11515
The main difficulty of person re-identification (ReID) lies in collecting annotated data and transferring the model across different domains. This paper presents UnrealPerson, a novel pipeline that makes full use of unreal image data to dec

rease the costs in both the training and deployment stages. Its fundamental part is a system that can generate synthesized images of high-quality and from controllable distributions. Instance-level annotation goes with the synthesized data and is almost free. We point out some details in image synthesis that largely impact the data quality. With 3,000 IDs and 120,000 instances, our method achieves a 38.5% rank-1 accuracy when being directly transferred to MSMT17. It almost doubles the former record using synthesized data and even surpasses previous direct transfer records using real data. This offers a good basis for unsupervised domain adaption, where our pre-trained model is easily plugged into the state-of-the-art algorithms towards higher accuracy. In addition, the data distribution can be flexibly adjusted to fit some corner ReID scenarios, which widens the application of our pipeline. We publish our data synthesis toolkit and synthesized data in https://github.com/FlyHighest/UnrealPerson.
****
Iterative Filter Adaptive Network for Single Image Defocus Deblurring
Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, Seungyong Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2034-2042
We propose a novel end-to-end learning-based approach for single image defocus deblurring. The proposed approach is equipped with a novel Iterative Filter Adaptive Network (IFAN) that is specifically designed to handle spatially-varying and large defocus blur. For adaptively handling spatially-varying blur, IFAN predicts pixel-wise deblurring filters, which are applied to defocused features of an input image to generate deblurred features. For effectively managing large blur, IFAN models deblurring filters as stacks of small-sized separable filters. Predicted separable deblurring filters are applied to defocused features using a novel Iterative Adaptive Convolution (IAC) layer. We also propose a training scheme based on defocus disparity estimation and reblurring, which significantly boosts the deblurring quality. We demonstrate that our method achieves state-of-the-art performance both quantitatively and qualitatively on real-world images.
****
UPFlow: Upsampling Pyramid for Unsupervised Optical Flow Learning
Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1045-1054
We present an unsupervised learning approach for optical flow estimation by improving the upsampling and learning of pyramid network. We design a self-guided upsample module to tackle the interpolation blur problem caused by bilinear upsampling between pyramid levels. Moreover, we propose a pyramid distillation loss to add supervision for intermediate levels via distilling the finest flow as pseudo labels. By integrating these two components together, our method achieves the best performance for unsupervised optical flow learning on multiple leading benchmarks, including MPI-SIntel, KITTI 2012 and KITTI 2015. In particular, we achieve EPE=1.4 on KITTI 2012 and F1=9.38% on KITTI 2015, which outperform the previous state-of-the-art methods by 22.2% and 15.7%, respectively.
****
House-GAN++: Generative Adversarial Layout Refinement Network towards Intelligent Computational Agent for Professional Architects
Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13632-13641
This paper proposes a generative adversarial layout refinement network for automated floorplan generation. Our architecture is an integration of a graph-constrained relational GAN and a conditional GAN, where a previously generated layout becomes the next input constraint, enabling iterative refinement. A surprising discovery of our research is that a simple non-iterative training process, dubbed component-wise GT-conditioning, is effective in learning such a generator. The iterative generator further allows us to improve a metric of choice via meta-optimization techniques by controlling when to pass which input constraints during iterative refinement. Our qualitative and quantitative evaluation based on the th

ree standard metrics demonstrate that the proposed system makes significant impr ovements over the current state-of-the-art, even competitive against the ground-truth floorplans, designed by professional architects. Code, model, and data are available at https://ennauata.github.io/houseganpp/page.html.
****

HDR Environment Map Estimation for Real-Time Augmented Reality
Gowri Somanath, Daniel Kurz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11298-11306
We present a method to estimate an HDR environment map from a narrow field-of-vi ew LDR camera image in real-time. This enables perceptually appealing reflection s and shading on virtual objects of any material finish, from mirror to diffuse, rendered into a real environment using augmented reality. Our method is based o n our efficient convolutional neural network, EnvMapNet, trained end-to-end with two novel losses, ProjectionLoss for the generated image, and ClusterLoss for a dversarial training. Through qualitative and quantitative comparison to state-of -the-art methods, we demonstrate that our algorithm reduces the directional erro r of estimated light sources by more than 50%, and achieves 3.7 times lower Frec het Inception Distance (FID). We further showcase a mobile application that is a ble to run our neural network model in under 9ms on an iPhone XS, and render in real-time, visually coherent virtual objects in previously unseen real-world env ironments.
****

OTA: Optimal Transport Assignment for Object Detection
Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, Jian Sun; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 30 3-312
Recent advances in label assignment in object detection mainly seek to independe ntly define positive/negative training samples for each ground-truth (gt) object . In this paper, we innovatively revisit the label assignment from a global pers pective and propose to formulate the assigning procedure as an Optimal Transport (OT) problem -- a well-studied topic in Optimization Theory. Concretely, we def ine the unit transportation cost between each demander (anchor) and supplier (gt) pair as the weighted summation of their classification and regression losses. After formulation, finding the best assignment solution is converted to solve th e optimal transport plan at minimal transportation costs, which can be solved vi a Sinkhorn-Knopp Iteration. On COCO, a single FCOS-ResNet-50 detector equipped w ith Optimal Transport Assignment (OTA) can reach 40.7% mAP under 1x scheduler, o utperforming all other existing assigning methods. Extensive experiments conduct ed on COCO and CrowdHuman further validate the effectiveness of our proposed OTA , especially its superiority in crowd scenarios. The code is available at https://github.com/Megvii-BaseDetection/OTA.
****

Progressive Semantic Segmentation
Chuong Huynh, Anh Tuan Tran, Khoa Luu, Minh Hoai; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16755-1676 4
The objective of this work is to segment high-resolution images without overload ing GPU memory usage or losing the fine details in the output segmentation map. The memory constraint means that we must either downsample the big image or divi de the image into local patches for separate processing. However, the former app roach would lose the fine details, while the latter can be ambiguous due to the lack of a global picture. In this work, we present MagNet, a multi-scale framewo rk that resolves local ambiguity by looking at the image at multiple magnificati on levels. MagNet has multiple processing stages, where each stage corresponds t o a magnification level, and the output of one stage is fed into the next stage for coarse-to-fine information propagation. Each stage analyzes the image at a h igher resolution than the previous stage, recovering the previously lost details due to the lossy downsampling step, and the segmentation output is progressivel y refined through the processing stages. Experiments on three high-resolution da tasets of urban views, aerial scenes, and medical images shows that MagNet consi

stently outperforms the state-of-the-art methods by a significant margin.
****
BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyo
nd
Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, Chen Change Loy; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021
, pp. 4947-4956
Video super-resolution (VSR) approaches tend to have more components than the im
age counterparts as they need to exploit the additional temporal dimension. Comp
lex designs are not uncommon. In this study, we wish to untangle the knots and r
econsider some most essential components for VSR guided by four basic functional
ities, i.e., Propagation, Alignment, Aggregation, and Upsampling. By reusing som
e existing components added with minimal redesigns, we show a succinct pipeline,
 BasicVSR, that achieves appealing improvements in terms of speed and restoratio
n quality in comparison to many state-of-the-art algorithms. We conduct systemat
ic analysis to explain how such gain can be obtained and discuss the pitfalls. W
e further show the extensibility of BasicVSR by presenting an information-refill
 mechanism and a coupled propagation scheme to facilitate information aggregatio
n. The BasicVSR and its extension, IconVSR, can serve as strong baselines for fu
ture VSR approaches.
****
Efficient Multi-Stage Video Denoising With Recurrent Spatio-Temporal Fusion
Matteo Maggioni, Yibin Huang, Cheng Li, Shuai Xiao, Zhongqian Fu, Fenglong Song;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2021, pp. 3466-3475
In recent years, denoising methods based on deep learning have achieved unparall
eled performance at the cost of large computational complexity. In this work, we
 propose an Efficient Multi-stage Video Denoising algorithm, called EMVD, to dra
stically reduce the complexity while maintaining or even improving the performan
ce. First, a fusion stage reduces the noise through a recursive combination of a
ll past frames in the video. Then, a denoising stage removes the noise in the fu
sed frame. Finally, a refinement stage restores the missing high frequency in th
e denoised frame. All stages operate on a transform-domain representation obtain
ed by learnable and invertible linear operators which simultaneously increase ac
curacy and decrease complexity of the model. A single loss on the final output i
s sufficient for successful convergence, hence making EMVD easy to train. Experi
ments on real raw data demonstrate that EMVD outperforms the state of the art wh
en complexity is constrained, and even remains competitive against methods whose
 complexities are several orders of magnitude higher. Further, the low complexit
y and memory requirements of EMVD enable real-time video denoising on commercial
 SoC in mobile devices.
****
Self-Supervised Simultaneous Multi-Step Prediction of Road Dynamics and Cost Map
Elmira Amirloo, Mohsen Rohani, Ershad Banijamali, Jun Luo, Pascal Poupart; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2021, pp. 8494-8503
In this paper we propose a system consisting of a modular network and a trajecto
ry planner. The network simultaneously predicts Occupancy Grid Maps (OGMs) and e
stimates space-time cost maps (CMs) corresponding to the areas around the vehicl
e. The trajectory planner computes the cost of a set of predefined trajectories
and chooses the one with the lowest cost. Training this network is done in a sel
f-supervised manner which desirably do not require any labeled data. The propose
d training objective takes into account the accuracy of OGM predictions as well
as contextual information and human driver behavior. Training these modules end-
to-end makes each module aware of the errors caused by the other components of t
he system. We show that our proposed method can lead to the selection of low cos
t trajectories with a low collision rate and road violation in fairly long plann
ing horizons.
****
Probabilistic Tracklet Scoring and Inpainting for Multiple Object Tracking

Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14329-14339
Despite the recent advances in multiple object tracking (MOT), achieved by joint detection and tracking, dealing with long occlusions remains a challenge. This is due to the fact that such techniques tend to ignore the long-term motion information. In this paper, we introduce a probabilistic autoregressive motion model to score tracklet proposals by directly measuring their likelihood. This is achieved by training our model to learn the underlying distribution of natural tracklets. As such, our model allows us not only to assign new detections to existing tracklets, but also to inpaint a tracklet when an object has been lost for a long time, e.g., due to occlusion, by sampling tracklets so as to fill the gap caused by misdetections. Our experiments demonstrate the superiority of our approach at tracking objects in challenging sequences; it outperforms the state of the art in most standard MOT metrics on multiple MOT benchmark datasets, including MOT16, MOT17, and MOT20.
****
Stay Positive: Non-Negative Image Synthesis for Augmented Reality
Katie Luo, Guandao Yang, Wenqi Xian, Harald Haraldsson, Bharath Hariharan, Serge Belongie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10050-10060
In applications such as optical see-through and projector augmented reality, producing images amounts to solving non-negative image generation, where one can only add light to an existing image. Most image generation methods, however, are ill-suited to this problem setting, as they make the assumption that one can assign arbitrary color to each pixel. In fact, naive application of existing methods fails even in simple domains such as MNIST digits, since one cannot create darker pixels by adding light. We know, however, that the human visual system can be fooled by optical illusions involving certain spatial configurations of brightness and contrast. Our key insight is that one can leverage this behavior to produce high quality images with negligible artifacts. For example, we can create the illusion of darker patches by brightening surrounding pixels. We propose a novel optimization procedure to produce images that satisfy both semantic and non-negativity constraints. Our approach can incorporate existing state-of-the-art methods, and exhibits strong performance in a variety of tasks including image-to-image translation and style transfer.
****
3D-to-2D Distillation for Indoor Scene Parsing
Zhengzhe Liu, Xiaojuan Qi, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4464-4474
Indoor scene semantic parsing from RGB images is very challenging due to occlusions, object distortion, and viewpoint variations. Going beyond prior works that leverage geometry information, typically paired depth maps, we present a new approach, a 3D-to-2D distillation framework, that enables us to leverage 3D features extracted from large-scale 3D data repository (e.g., ScanNet-v2) to enhance 2D features extracted from RGB images. Our work has three novel contributions. First, we distill 3D knowledge from a pretrained 3D network to supervise a 2D network to learn simulated 3D features from 2D features during the training, so the 2D network can infer without requiring 3D data. Second, we design a two-stage dimension normalization scheme to calibrate the 2D and 3D features for better integration. Third, we design a semantic-aware adversarial training model to extend our framework for training with unpaired 3D data. Extensive experiments on various datasets, ScanNet-V2, S3DIS, and NYU-v2, demonstrate the superiority of our approach. Also, experimental results show that our 3D-to-2D distillation improves the model generalization.
****
Learning the Best Pooling Strategy for Visual Semantic Embedding
Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, Changhu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15789-15798

Visual Semantic Embedding (VSE) is a dominant approach for vision-language retrieval, which aims at learning a deep embedding space such that visual data are embedded close to their semantic text labels or descriptions. Recent VSE models use complex methods to better contextualize and aggregate multi-modal features into holistic embeddings. However, we discover that surprisingly simple (but carefully selected) global pooling functions (e.g., max pooling) outperform those complex models, across different feature extractors. Despite its simplicity and effectiveness, seeking the best pooling function for different data modality and feature extractor is costly and tedious, especially when the size of features varies (e.g., text, video). Therefore, we propose a Generalized Pooling Operator (GPO), which learns to automatically adapt itself to the best pooling strategy for different features, requiring no manual tuning while staying effective and efficient. We extend the VSE model using this proposed GPO and denote it as VSE. Without bells and whistles, VSE outperforms previous VSE methods significantly on image-text retrieval benchmarks across popular feature extractors. With a simple adaptation, variants of VSE further demonstrate its strength by achieving the new state of the art on two video-text retrieval datasets. Comprehensive experiments and visualizations confirm that GPO always discovers the best pooling strategy and can be a plug-and-play feature aggregation module for standard VSE models.
****

GLAVNet: Global-Local Audio-Visual Cues for Fine-Grained Material Recognition
Fengmin Shi, Jie Guo, Haonan Zhang, Shan Yang, Xiying Wang, Yanwen Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14433-14442
In this paper, we aim to recognize materials with combined use of auditory and visual perception. To this end, we construct a new dataset named GLAudio that consists of both the geometry of the object being struck and the sound captured from either modal sound synthesis (for virtual objects) or real measurements (for real objects). Besides global geometries, our dataset also takes local geometries around different hitpoints into consideration. This local information is less explored in existing datasets. We demonstrate that local geometry has a greater impact on the sound than the global geometry and offers more cues in material recognition. To extract features from different modalities and perform proper fusion, we propose a new deep neural network GLAVNet that comprises multiple branches and a well-designed fusion module. Once trained on GLAudio, our GLAVNet provides state-of-the-art performance on material identification and supports fine-grained material categorization.
****

Refining Pseudo Labels With Clustering Consensus Over Generations for Unsupervised Object Re-Identification
Xiao Zhang, Yixiao Ge, Yu Qiao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3436-3445
Unsupervised object re-identification targets at learning discriminative representations for object retrieval without any annotations. Clustering-based methods conduct training with the generated pseudo labels and currently dominate this research direction. However, they still suffer from the issue of pseudo label noise. To tackle the challenge, we propose to properly estimate pseudo label similarities between consecutive training generations with clustering consensus and refine pseudo labels with temporally propagated and ensembled pseudo labels. To the best of our knowledge, this is the first attempt to leverage the spirit of temporal ensembling to improve classification with dynamically changing classes over generations. The proposed pseudo label refinery strategy is simple yet effective and can be seamlessly integrated into existing clustering-based unsupervised re-identification methods. With our proposed approach, state-of-the-art method can be further boosted with up to 8.8% mAP improvements on the challenging MSMT17 dataset.
****

Regularizing Generative Adversarial Networks Under Limited Data
Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, Weilong Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,

pp. 7921-7931
Recent years have witnessed the rapid progress of generative adversarial network s (GANs). However, the success of the GAN models hinges on a large amount of tra ining data. This work proposes a regularization approach for training robust GAN models on limited data. We theoretically show a connection between the regulari zed loss and an f-divergence called LeCam-divergence, which we find is more robu st under limited training data. Extensive experiments on several benchmark datas ets demonstrate that the proposed regularization scheme 1) improves the generali zation performance and stabilizes the learning dynamics of GAN models under limi ted training data, and 2) complements the recent data augmentation methods. Thes e properties facilitate training GAN models to achieve state-of-the-art performa nce when only limited training data of the ImageNet benchmark is available. The source code is available at https://github.com/google/lecam-gan.
****
Skeleton Merger: An Unsupervised Aligned Keypoint Detector
Ruoxi Shi, Zhengrong Xue, Yang You, Cewu Lu; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 43-52
Detecting aligned 3D keypoints is essential under many scenarios such as object tracking, shape retrieval and robotics. However, it is generally hard to prepare a high-quality dataset for all types of objects due to the ambiguity of keypoin t itself. Meanwhile, current unsupervised detectors are unable to generate align ed keypoints with good coverage. In this paper, we propose an unsupervised align ed keypoint detector, Skeleton Merger, which utilizes skeletons to reconstruct o bjects. It is based on an Autoencoder architecture. The encoder proposes keypoin ts and predicts activation strengths of edges between keypoints. The decoder per forms uniform sampling on the skeleton and refines it into small point clouds wi th pointwise offsets. Then the activation strengths are applied and the sub-clou ds are merged. Composite Chamfer Distance (CCD) is proposed as a distance betwee n the input point cloud and the reconstruction composed of sub-clouds masked by activation strengths. We demonstrate that Skeleton Merger is capable of detectin g semantically-rich salient keypoints with good alignment, and shows comparable performance to supervised methods on the KeypointNet dataset. It is also shown t hat the detector is robust to noise and subsampling. Our code is available at ht tps://github.com/eliphatfs/SkeletonMerger.
****
Regularizing Neural Networks via Adversarial Model Perturbation
Yaowei Zheng, Richong Zhang, Yongyi Mao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8156-8165
Effective regularization techniques are highly desired in deep learning for alle viating overfitting and improving generalization. This work proposes a new regul arization scheme, based on the understanding that the flat local minima of the e mpirical risk cause the model to generalize better. This scheme is referred to a s adversarial model perturbation (AMP), where instead of directly minimizing the empirical risk, an alternative "AMP loss" is minimized via SGD. Specifically, t he AMP loss is obtained from the empirical risk by applying the "worst" norm-bou nded perturbation on each point in the parameter space. Comparing with most exis ting regularization schemes, AMP has strong theoretical justifications, in that minimizing the AMP loss can be shown theoretically to favour flat local minima o f the empirical risk. Extensive experiments on various modern deep architectures establish AMP as a new state of the art among regularization schemes. Our code is available at https://github.com/hiyouga/AMP-Regularizer.
****
Learning by Aligning Videos in Time
Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N. Syed, Andrey Konin, Zee shan Zia, Quoc-Huy Tran; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2021, pp. 5548-5558
We present a self-supervised approach for learning video representations using t emporal video alignment as a pretext task, while exploiting both frame-level and video-level information. We leverage a novel combination of temporal alignment loss and temporal regularization terms, which can be used as supervision signals

for training an encoder network. Specifically, the temporal alignment loss (i.e., Soft-DTW) aims for the minimum cost for temporally aligning videos in the embedding space. However, optimizing solely for this term leads to trivial solutions, particularly, one where all frames get mapped to a small cluster in the embedding space. To overcome this problem, we propose a temporal regularization term (i.e., Contrastive-IDM) which encourages different frames to be mapped to different points in the embedding space. Extensive evaluations on various tasks, including action phase classification, action phase progression, and fine-grained frame retrieval, on three datasets, namely Pouring, Penn Action, and IKEA ASM, show superior performance of our approach over state-of-the-art methods for self-supervised representation learning from videos. In addition, our method provides significant performance gain where labeled data is lacking.
****
Contrastive Neural Architecture Search With Neural Architecture Comparators
Yaofo Chen, Yong Guo, Qi Chen, Minli Li, Wei Zeng, Yaowei Wang, Mingkui Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9502-9511
One of the key steps in Neural Architecture Search (NAS) is to estimate the performance of candidate architectures. Existing methods either directly use the validation performance or learn a predictor to estimate the performance. However, these methods can be either computationally expensive or very inaccurate, which may severely affect the search efficiency and performance. Moreover, as it is very difficult to annotate architectures with accurate performance on specific tasks, learning a promising performance predictor is often non-trivial due to the lack of labeled data. In this paper, we argue that it may not be necessary to estimate the absolute performance for NAS. On the contrary, we may need only to understand whether an architecture is better than a baseline one. However, how to exploit this comparison information as the reward and how to well use the limited labeled data remains two great challenges. In this paper, we propose a novel Contrastive Neural Architecture Search (CTNAS) method which performs architecture search by taking the comparison results between architectures as the reward. Specifically, we design and learn a Neural Architecture Comparator (NAC) to compute the probability of candidate architectures being better than a baseline one. Moreover, we present a baseline updating scheme to improve the baseline iteratively in a curriculum learning manner. More critically, we theoretically show that learning NAC is equivalent to optimizing the ranking over architectures. Extensive experiments in three search spaces demonstrate the superiority of our CTNAS over existing methods.
****
Implicit Feature Alignment: Learn To Convert Text Recognizer to Text Spotter
Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Dezhi Peng, Zhe Li, Mengchao He, Yongpan Wang, Canjie Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5973-5982
Text recognition is a popular research subject with many associated challenges. Despite the considerable progress made in recent years, the text recognition task itself is still constrained to solve the problem of reading cropped line text images and serves as a subtask of optical character recognition (OCR) systems. As a result, the final text recognition result is limited by the performance of the text detector. In this paper, we propose a simple, elegant and effective paradigm called Implicit Feature Alignment (IFA), which can be easily integrated into current text recognizers, resulting in a novel inference mechanism called IFA-inference. This enables an ordinary text recognizer to process multi-line text such that text detection can be completely freed. Specifically, we integrate IFA into the two most prevailing text recognition streams (attention-based and CTC-based) and propose attention-guided dense prediction (ADP) and Extended CTC (ExCTC). Furthermore, the Wasserstein-based Hollow Aggregation Cross-Entropy (WH-ACE) is proposed to suppress negative predictions to assist in training ADP and ExCTC. We experimentally demonstrate that IFA achieves state-of-the-art performance on end-to-end document recognition tasks while maintaining the fastest speed, and ADP and ExCTC complement each other on the perspective of different applicatio

n scenarios. Code will be available at https://github.com/Wang-Tianwei/Implicit-feature-alignment.
****
Populating 3D Scenes by Learning Human-Scene Interaction
Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14708-14718
Humans live within a 3D space and constantly interact with it to perform tasks. Such interactions involve physical contact between surfaces that is semantically meaningful. Our goal is to learn how humans interact with scenes and leverage this to enable virtual characters to do the same. To that end, we introduce a novel Human-Scene Interaction (HSI) model that encodes proximal relationships, called POSA for "Pose with prOximitieS and contActs". The representation of interaction is body-centric, which enables it to generalize to new scenes. Specifically, POSA augments the SMPL-X parametric human body model such that, for every mesh vertex, it encodes (a) the contact probability with the scene surface and (b) the corresponding semantic scene label. We learn POSA with a VAE conditioned on the SMPL-X vertices, and train on the PROX dataset, which contains SMPL-X meshes of people interacting with 3D scenes, and the corresponding scene semantics from the PROX-E dataset. We demonstrate the value of POSA with two applications. First, we automatically place 3D scans of people in scenes. We use a SMPL-X model fit to the scan as a proxy and then find its most likely placement in 3D. POSA provides an effective representation to search for "affordances" in the scene that match the likely contact relationships for that pose. We perform a perceptual study that shows significant improvement over the state of the art on this task. Second, we show that POSA's learned representation of body-scene interaction supports monocular human pose estimation that is consistent with a 3D scene, improving on the state of the art. Our model and code are available for research purposes at https://posa.is.tue.mpg.de.
****
Variational Pedestrian Detection
Yuang Zhang, Huanyu He, Jianguo Li, Yuxi Li, John See, Weiyao Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11622-11631
Pedestrian detection in a crowd is a challenging task due to a high number of mutually-occluding human instances, which brings ambiguity and optimization difficulties to the current IoU-based ground truth assignment procedure in classical object detection methods. In this paper, we develop a unique perspective of pedestrian detection as a variational inference problem. We formulate a novel and efficient algorithm for pedestrian detection by modeling the dense proposals as a latent variable while proposing a customized Auto-Encoding Variational Bayes (AEVB) algorithm. Through the optimization of our proposed algorithm, a classical detector can be fashioned into a variational pedestrian detector. Experiments conducted on CrowdHuman and CityPersons datasets show that the proposed algorithm serves as an efficient solution to handle the dense pedestrian detection problem for the case of single-stage detectors. Our method can also be flexibly applied to two-stage detectors, achieving notable performance enhancement.
****
SIPSA-Net: Shift-Invariant Pan Sharpening With Moving Object Alignment for Satellite Imagery
Jaehyup Lee, Soomin Seo, Munchurl Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10166-10174
Pan-sharpening is a process of merging a high-resolution (HR) panchromatic (PAN) image and its corresponding low-resolution (LR) multi-spectral (MS) image to create an HR-MS and pan-sharpened image. However, due to the different sensors' locations, characteristics and acquisition time, PAN and MS image pairs often tend to have various amounts of misalignment. Conventional deep-learning-based methods that were trained with such misaligned PAN-MS image pairs suffer from diverse artifacts such as double-edge and blur artifacts in the resultant PAN-sharpened images. In this paper, we propose a novel framework called shift-invariant pan-

sharpening with moving object alignment (SIPSA-Net) which is the first method to take into account such large misalignment of moving object regions for PAN sharpening. The SISPA-Net has a feature alignment module (FAM) that can adjust one feature to be aligned to another feature, even between the two different PAN and MS domains. For better alignment in pan-sharpened images, a shift-invariant spectral loss is newly designed, which ignores the inherent misalignment in the original MS input, thereby having the same effect as optimizing the spectral loss with a well-aligned MS image. Extensive experimental results show that our SIPSA-Net can generate pan-sharpened images with remarkable improvements in terms of visual quality and alignment, compared to the state-of-the-art methods.
****
Large-Scale Localization Datasets in Crowded Indoor Spaces
Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guerin, Gabriela Csurka, Martin Humenberger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3227-3236
Estimating the precise location of a camera using visual localization enables interesting applications such as augmented reality or robot navigation. This is particularly useful in indoor environments where other localization technologies, such as GNSS, fail. Indoor spaces impose interesting challenges on visual localization algorithms: occlusions due to people, textureless surfaces, large viewpoint changes, low light, repetitive textures, etc. Existing indoor datasets are either comparably small or do only cover a subset of the mentioned challenges. In this paper, we introduce 5 new indoor datasets for visual localization in challenging real-world environments. They were captured in a large shopping mall and a large metro station in Seoul, South Korea, using a dedicated mapping platform consisting of 10 cameras and 2 laser scanners. In order to obtain accurate ground truth camera poses, we developed a robust LiDAR SLAM which provides initial poses that are then refined using a novel structure-from-motion based optimization. We present a benchmark of modern visual localization algorithms on these challenging datasets showing superior performance of structure-based methods using robust image features. The datasets are available at: https://naverlabs.com/datasets
****
Distilling Causal Effect of Data in Class-Incremental Learning
Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3957-3966
We propose a causal framework to explain the catastrophic forgetting in Class-Incremental Learning (CIL) and then derive a novel distillation method that is orthogonal to the existing anti-forgetting techniques, such as data replay and feature/label distillation. We first 1) place CIL into the framework, 2) answer why the forgetting happens: the causal effect of the old data is lost in new training, and then 3) explain how the existing techniques mitigate it: they bring the causal effect back. Based on the causal framework, we propose to distill the Colliding Effect between the old and the new data, which is fundamentally equivalent to the causal effect of data replay, but without any cost of replay storage. Thanks to the causal effect analysis, we can further capture the Incremental Momentum Effect of the data stream, removing which can help to retain the old effect overwhelmed by the new data effect, and thus alleviate the forgetting of the old class in testing. Extensive experiments on three CIL benchmarks: CIFAR-100, ImageNet-Sub&Full, show that the proposed causal effect distillation can improve various state-of-the-art CIL methods by a large margin (0.72%-9.06%)
****
Backdoor Attacks Against Deep Learning Systems in the Physical World
Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, Ben Y. Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6206-6215
Backdoor attacks embed hidden malicious behaviors into deep learning models, which only activate and cause misclassifications on model inputs containing a speci

fic "trigger." Existing works on backdoor attacks and defenses, however, mostly focus on digital attacks that apply digitally generated patterns as triggers. A critical question remains unanswered: "can backdoor attacks succeed using physical objects as triggers, making them a credible threat against deep learning systems in the real world?" We conduct a detailed empirical study to explore this question for facial recognition, a critical deep learning task. Using 7 physical objects as triggers, we collect a custom dataset of 3205 images of 10 volunteers and use it to study the feasibility of "physical" backdoor attacks under a variety of real-world conditions. Our study reveals two key findings. First, physical backdoor attacks can be highly successful if they are carefully configured to overcome the constraints imposed by physical objects. In particular, the placement of successful triggers is largely constrained by the victim model's dependence on key facial features. Second, four of today's state-of-the-art defenses against (digital) backdoors are ineffective against physical backdoors, because the use of physical objects breaks core assumptions used to construct these defenses. Our study confirms that (physical) backdoor attacks are not a hypothetical phenomenon but rather pose a serious real-world threat to critical classification tasks. We need new and more robust defenses against backdoors in the physical world.
****
A Multiplexed Network for End-to-End, Multilingual OCR
Jing Huang, Guan Pang, Rama Kovvuri, Mandy Toh, Kevin J Liang, Praveen Krishnan, Xi Yin, Tal Hassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4547-4557
Recent advances in OCR have shown that an end-to-end (E2E) training pipeline that includes both detection and recognition leads to the best results. However, many existing methods focus primarily on Latin-alphabet languages, often even only case-insensitive English characters. In this paper, we propose an E2E approach, Multiplexed Multilingual Mask TextSpotter, that performs script identification at the word level and handles different scripts with different recognition heads, all while maintaining a unified loss that simultaneously optimizes script identification and multiple recognition heads. Experiments show that our method outperforms single-head model with similar parameters in end-to-end recognition tasks, and achieves state-of-the-art results on MLT17 and MLT19 joint text detection and script identification benchmarks. We believe that our work is a step towards end-to-end trainable and scalable multilingual multi-purpose OCR system.
****
Semi-Supervised Semantic Segmentation With Directional Context-Aware Consistency
Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1205-1214
Semantic segmentation has made tremendous progress in recent years. However, satisfying performance highly depends on a large number of pixel-level annotations. Therefore, in this paper, we focus on the semi-supervised segmentation problem where only a small set of labeled data is provided with a much larger collection of totally unlabeled images. Nevertheless, due to the limited annotations, models may overly rely on the contexts available in the training data, which causes poor generalization to the scenes unseen before. A preferred high-level representation should capture the contextual information while not losing self-awareness. Therefore, we propose to maintain the context-aware consistency between features of the same identity but with different contexts, making the representations robust to the varying environments. Moreover, we present the Directional Contrastive Loss (DC Loss) to accomplish the consistency in a pixel-to-pixel manner, only requiring the feature with lower quality to be aligned towards its counterpart. In addition, to avoid the false-negative samples and filter the uncertain positive samples, we put forward two sampling strategies. Extensive experiments show that our simple yet effective method surpasses current state-of-the-art methods by a large margin and also generalizes well with extra image-level annotations.
****

Causal Hidden Markov Model for Time Series Disease Forecasting
Jing Li, Botong Wu, Xinwei Sun, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12105-12114
We propose a causal hidden Markov model to achieve robust prediction of irreversible disease at an early stage, which is safety-critical and vital for medical treatment in early stages. Specifically, we introduce the hidden variables which propagate to generate medical data at each time step. To avoid learning spurious correlation (e.g., confounding bias), we explicitly separate these hidden variables into three parts: a) the disease (clinical)-related part; b) the disease (non-clinical)-related part; c) others, with only a),b) causally related to the disease however c) may contain spurious correlations (with the disease) inherited from the data provided. With personal attributes and disease label respectively provided as side information and supervision, we prove that these disease-related hidden variables can be disentangled from others, implying the avoidance of spurious correlation for generalization to medical data from other (out-of-) distributions. Guaranteed by this result, we propose a sequential variational auto-encoder with a reformulated objective function. We apply our model to the early prediction of peripapillary atrophy and achieve promising results on out-of-distribution test data. Further, the ablation study empirically shows the effectiveness of each component in our method. And the visualization shows the accurate identification of lesion regions from others.
****
Generalizable Pedestrian Detection: The Elephant in the Room
Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11328-11337
Pedestrian detection is used in many vision based applications ranging from video surveillance to autonomous driving. Despite achieving high performance, it is still largely unknown how well existing detectors generalize to unseen data. This is important because a practical detector should be ready to use in various scenarios in applications. To this end, we conduct a comprehensive study in this paper, using a general principle of direct cross-dataset evaluation. Through this study, we find that existing state-of-the-art pedestrian detectors, though perform quite well when trained and tested on the same dataset, generalize poorly in cross dataset evaluation. We demonstrate that there are two reasons for this trend. Firstly, their designs (e.g. anchor settings) may be biased towards popular benchmarks in the traditional single-dataset training and test pipeline, but as a result largely limit their generalization capability. Secondly, the training source is generally not dense in pedestrians and diverse in scenarios. Under direct cross-dataset evaluation, surprisingly, we find that a general purpose object detector, without pedestrian-tailored adaptation in design, generalizes much better compared to existing state-of-the-art pedestrian detectors. Furthermore, we illustrate that diverse and dense datasets, collected by crawling the web, serve to be an efficient source of pre-training for pedestrian detection. Accordingly, we propose a progressive training pipeline and find that it works well for autonomous-driving oriented pedestrian detection. Consequently, the study conducted in this paper suggests that more emphasis should be put on cross-dataset evaluation for the future design of generalizable pedestrian detectors. Code and models can be accessed at https://github.com/hasanirtiza/Pedestron.
****
Focus on Local: Detecting Lane Marker From Bottom Up via Key Point
Zhan Qu, Huan Jin, Yang Zhou, Zhen Yang, Wei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14122-14130
Mainstream lane marker detection methods are implemented by predicting the overall structure and deriving parametric curves through post-processing. Complex lane line shapes require high-dimensional output of CNNs to model global structures, which further increases the demand for model capacity and training data. In contrast, the locality of a lane marker has finite geometric variations and spatial coverage. We propose a novel lane marker detection solution, FOLOLane, that fo

cuses on modeling local patterns and achieving prediction of global structures in a bottom-up manner. Specifically, the CNN models low-complexity local patterns with two separate heads, the first one predicts the existence of key points, and the second refines the location of key points in the local range and correlates key points of the same lane line. The locality of the task is consistent with the limited FOV of the feature in CNN, which in turn leads to more stable training and better generalization. In addition, an efficiency-oriented decoding algorithm was proposed as well as a greedy one, which achieving 36% runtime gains at the cost of negligible performance degradation. Both of the two decoders integrated local information into the global geometry of lane markers. In the absence of a complex network architecture design, the proposed method greatly outperforms all existing methods on public datasets while achieving the best state-of-the-art results and real-time processing simultaneously.
****
Memory-Guided Unsupervised Image-to-Image Translation
Somi Jeong, Youngjung Kim, Eungbean Lee, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6558-6567
We present a novel unsupervised framework for instance-level image-to-image translation. Although recent advances have been made by incorporating additional object annotations, existing methods often fail to handle images with multiple disparate objects. The main cause is that, during inference, they apply a global style to the whole image and do not consider the large style discrepancy between instance and background, or within instances. To address this problem, we propose a class-aware memory network that explicitly reasons about local style variations. A key-values memory structure, with a set of read/update operations, is introduced to record class-wise style variations and access them without requiring an object detector at the test time. The key stores a domain-agnostic content representation for allocating memory items, while the values encode domain-specific style representations. We also present a feature contrastive loss to boost the discriminative power of memory items. We show that by incorporating our memory, we can transfer class-aware and accurate style representations across domains. Experimental results demonstrate that our model outperforms recent instance-level methods and achieves state-of-the-art performance.
****
Incremental Few-Shot Instance Segmentation
Dan Andrei Ganea, Bas Boom, Ronald Poppe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1185-1194
Few-shot instance segmentation methods are promising when labeled training data for novel classes is scarce. However, current approaches do not facilitate flexible addition of novel classes. They also require that examples of each class are provided at train and test time, which is memory intensive. In this paper, we address these limitations by presenting the first incremental approach to few-shot instance segmentation: iMTFA. We learn discriminative embeddings for object instances that are merged into class representatives. Storing embedding vectors rather than images effectively solves the memory overhead problem. We match these class embeddings at the RoI-level using cosine similarity. This allows us to add new classes without the need for further training or access to previous training data. In a series of experiments, we consistently outperform the current state-of-the-art. Moreover, the reduced memory requirements allow us to evaluate, for the first time, few-shot instance segmentation performance on all classes in COCO jointly.
****
Mining Better Samples for Contrastive Learning of Temporal Correspondence
Sangryul Jeon, Dongbo Min, Seungryong Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1034-1044
We present a novel framework for contrastive learning of pixel-level representation using only unlabeled video. Without the need of ground-truth annotation, our method is capable of collecting well-defined positive correspondences by measur

ing their confidences and well-defined negative ones by appropriately adjusting their hardness during training. This allows us to suppress the adverse impact of ambiguous matches and prevent a trivial solution from being yielded by too hard or too easy negative samples. To accomplish this, we incorporate three different criteria that ranges from a pixel-level matching confidence to a video-level one into a bottom-up pipeline, and plan a curriculum that is aware of current representation power for the adaptive hardness of negative samples during training. With the proposed method, state-of-the-art performance is attained over the latest approaches on several video label propagation tasks.
****
Scene-Aware Generative Network for Human Motion Synthesis
Jingbo Wang, Sijie Yan, Bo Dai, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12206-12215
We revisit human motion synthesis, a task useful in various real-world applications, in this paper. Whereas a number of methods have been developed previously for this task, they are often limited in two aspects: 1) focus on the poses while leaving the location movement behind, and 2) ignore the impact of the environment on the human motion. In this paper, we propose a new framework, with the interaction between the scene and the human motion is taken into account. Considering the uncertainty of human motion, we formulate this task as a generative task, whose objective is to generate plausible human motion conditioned on both the scene and the human's initial position. This framework factorizes the distribution of human motions into a distribution of movement trajectories conditioned on scenes and that of body pose dynamics conditioned on both scenes and trajectories. We further derive a GAN-based learning approach, with discriminators to enforce the compatibility between the human motion and the contextual scene as well as the 3D-to-2D projection constraints. We assess the effectiveness of the proposed method on two challenging datasets, which cover both synthetic and real-world environmentemphasizes local structural constraints via depth-map crops, and a projection discriminator that emphasizes global structural constraints via 3D-to-2D motion projections. The effectiveness of our framework is comprehensively evaluated on two large challenging datasets, covering both a synthetic environment (GTA-IM) and a real environment (PROX)
****
Learning Neural Representation of Camera Pose with Matrix Representation of Pose Shift via View Synthesis
Yaxuan Zhu, Ruiqi Gao, Siyuan Huang, Song-Chun Zhu, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9959-9968
How to efficiently represent camera pose is an essential problem in 3D computer vision, especially in tasks like camera pose regression and novel view synthesis. Traditionally, 3D position of the camera is represented by Cartesian coordinate and the orientation is represented by Euler angle or quaternions. These representations are manually designed, which may not be the most efficient representation for downstream tasks. In this work, we propose an approach to learn neural representations of camera poses and 3D scenes, coupled with neural representations of local camera movements. Specifically, the camera pose and 3D scene are represented as vectors and the local camera movement is represented as a matrix operating on the vector of the camera pose. We demonstrate that the camera movement can further be parametrized as a matrix Lie algebra that underlies a rotation system in the neural space. The vector representations are then concatenated and generate the posed 2D image through a decoder network. The model is learned from only posed 2D images and corresponding camera poses, without access to depth or shape. We conduct extensive experiments on synthetic and real datasets. The results show that compared with other camera pose representations, our learned representation is more robust to noise in novel view synthesis and more effective in camera pose regression.
****
PML: Progressive Margin Loss for Long-Tailed Age Classification
Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, Xuehong Sun; Pro

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10503-10512

In this paper, we propose a progressive margin loss (PML) approach for unconstrained facial age classification. Conventional methods make strong assumption on that each class owns adequate instances to outline its data distribution, likely leading to bias prediction where the training samples are sparse across age classes. Instead, our PML aims to adaptively refine the age label pattern by enforcing a couple of margins, which fully takes in the in-between discrepancy of the intra-class variance, inter-class variance and class-center. Our PML typically incorporates with the ordinal margin and the variational margin, simultaneously plugging in the globally-tuned deep neural network paradigm. More specifically, the ordinal margin learns to exploit the correlated relationship of the real-world age labels. Accordingly, the variational margin is leveraged to minimize the influence of head classes that misleads the prediction of tailed samples. Moreover, our optimization carefully seeks a series of indicator curricula to achieve robust and efficient model training. Extensive experimental results on three face aging datasets demonstrate that our PML achieves compelling performance compared to state of the arts. Code will be made publicly.

****

Single Image Depth Prediction With Wavelet Decomposition

Michael Ramamonjisoa, Michael Firman, Jamie Watson, Vincent Lepetit, Daniyar Turmukhambetov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11089-11098

We present a novel method for predicting accurate depths from monocular images with high efficiency. This optimal efficiency is achieved by exploiting wavelet decomposition, which is integrated in a fully differentiable encoder-decoder architecture. We demonstrate that we can reconstruct high-fidelity depth maps by predicting sparse wavelet coefficients. In contrast with previous works, we show that wavelet coefficients can be learned without direct supervision on coefficients. Instead we supervise only the final depth image that is reconstructed through the inverse wavelet transform. We additionally show that wavelet coefficients can be learned in fully self-supervised scenarios, without access to ground-truth depth. Finally, we apply our method to different state-of-the-art monocular depth estimation models, in each case giving similar or better results compared to the original model, while requiring less than half the multiply-adds in the decoder network.

****

PVGNet: A Bottom-Up One-Stage 3D Object Detector With Integrated Multi-Level Features

Zhenwei Miao, Jikai Chen, Hongyu Pan, Ruiwen Zhang, Kaixuan Liu, Peihan Hao, Jun Zhu, Yang Wang, Xin Zhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3279-3288

Quantization-based methods are widely used in LiDAR points 3D object detection for its efficiency in extracting context information. Unlike image where the context information is distributed evenly over the object, most LiDAR points are distributed along the object boundary, which means the boundary features are more critical in LiDAR points 3D detection. However, quantization inevitably introduces ambiguity during both the training and inference stages. To alleviate this problem, we propose a one-stage and voting-based 3D detector, named Point-Voxel-Grid Network (PVGNet). In particular, PVGNet extracts point, voxel and grid-level features in a unified backbone architecture and produces point-wise fusion features. It segments LiDAR points into foreground and background, predicts a 3D bounding box for each foreground point, and performs group voting to get the final detection results. Moreover, we observe that instance-level point imbalance due to occlusion and observation distance also degrades the detection performance. A novel instance-aware focal loss is proposed to alleviate this problem and further improve the detection ability. We conduct experiments on the KITTI and Waymo datasets. Our proposed PVGNet outperforms previous state-of-the-art methods and ranks at the top of KITTI 3D/BEV detection leaderboards.

****

Exemplar-Based Open-Set Panoptic Segmentation Network
Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1175-1184
We extend panoptic segmentation to the open-world and introduce an open-set panoptic segmentation (OPS) task. The task requires to perform panoptic segmentation for not only known classes but also unknown ones that are not acknowledged during training. We investigate challenges of the task and present a benchmark dataset on top of an existing dataset, COCO. In addition, we propose a novel exemplar-based open-set panoptic segmentation network (EOPSN) inspired by exemplar theory. Our approach identifies a new class with exemplars, which constructs pseudo-ground-truths, based on clustering and augments the size of each class by adding new exemplars based on their similarity during training. We evaluate the proposed method on our benchmark and demonstrate the effectiveness of our proposals. The goal of our work is to draw the attention of the community to the recognition in open-world scenarios.
****
KOALAnet: Blind Super-Resolution Using Kernel-Oriented Adaptive Local Adjustment
Soo Ye Kim, Hyeonjun Sim, Munchurl Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10611-10620
Blind super-resolution (SR) methods aim to generate a high quality high resolution image from a low resolution image containing unknown degradations. However, natural images contain various types and amounts of blur: some may be due to the inherent degradation characteristics of the camera, but some may even be intentional, for aesthetic purposes (e.g. Bokeh effect). In the case of the latter, it becomes highly difficult for SR methods to disentangle the blur to remove, and that to leave as is. In this paper, we propose a novel blind SR framework based on kernel-oriented adaptive local adjustment (KOALA) of SR features, called KOALAnet, which jointly learns spatially-variant degradation and restoration kernels in order to adapt to the spatially-variant blur characteristics in real images. Our KOALAnet outperforms recent blind SR methods for synthesized LR images obtained with randomized degradations, and we further show that the proposed KOALAnet produces the most natural results for artistic photographs with intentional blur, which are not over-sharpened, by effectively handling images mixed with in-focus and out-of-focus areas.
****
Learning Deep Classifiers Consistent With Fine-Grained Novelty Detection
Jiacheng Cheng, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1664-1673
The problem of novelty detection in fine-grained visual classification (FGVC) is considered. An integrated understanding of the probabilistic and distance-based approaches to novelty detection is developed within the framework of convolutional neural networks (CNNs). It is shown that softmax CNN classifiers are inconsistent with novelty detection, because their learned class-conditional distributions and associated distance metrics are unidentifiable. A new regularization constraint, the class-conditional Gaussianity loss, is then proposed to eliminate this unidentifiability, and enforce Gaussian class-conditional distributions. This enables training Novelty Detection Consistent Classifiers (NDCCs) that are jointly optimal for classification and novelty detection. Empirical evaluations show that NDCCs achieve significant improvements over the state-of-the-art on both small- and large-scale FGVC datasets.
****
Multiple Object Tracking With Correlation Learning
Qiang Wang, Yun Zheng, Pan Pan, Yinghui Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3876-3886
Recent works have shown that convolutional networks have substantially improved the performance of multiple object tracking by simultaneously learning detection and appearance features. However, due to the local perception of the convolutional network structure itself, the long-range dependencies in both the spatial and temporal cannot be obtained efficiently. To incorporate the spatial layout, we

propose to exploit the local correlation module to model the topological relationship between targets and their surrounding environment, which can enhance the discriminative power of our model in crowded scenes. Specifically, we establish dense correspondences of each spatial location and its context, and explicitly constrain the correlation volumes through self-supervised learning. To exploit the temporal context, existing approaches generally utilize two or more adjacent frames to construct an enhanced feature representation, but the dynamic motion scene is inherently difficult to depict via CNNs. Instead, our paper proposes a learnable correlation operator to establish frame-to-frame matches over convolutional feature maps in the different layers to align and propagate temporal context. With extensive experimental results on the MOT datasets, our approach demonstrates the effectiveness of correlation learning with the superior performance and obtains state-of-the-art MOTA of 76.5% and IDF1 of 73.6% on MOT17.
****
SAIL-VOS 3D: A Synthetic Dataset and Baselines for Object Detection and 3D Mesh Reconstruction From Video Data
Yuan-Ting Hu, Jiahong Wang, Raymond A. Yeh, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1418-1428
Extracting detailed 3D information of objects from video data is an important goal for holistic scene understanding. While recent methods have shown impressive results when reconstructing meshes of objects from a single image, results often remain ambiguous as part of the object is unobserved. Moreover, existing image-based datasets for mesh reconstruction don't permit to study models which integrate temporal information. To alleviate both concerns we present SAIL-VOS 3D: a synthetic video dataset with frame-by-frame mesh annotations which extends SAIL-VOS. We also develop first baselines for reconstruction of 3D meshes from video data via temporal models. We demonstrate efficacy of the proposed baseline on SAIL-VOS 3D and Pix3D, showing that temporal information improves reconstruction quality. Resources and additional information are available at http://sailvos.web.illinois.edu.
****
PixMatch: Unsupervised Domain Adaptation via Pixelwise Consistency Training
Luke Melas-Kyriazi, Arjun K. Manrai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12435-12445
Unsupervised domain adaptation is a promising technique for semantic segmentation and other computer vision tasks for which large-scale data annotation is costly and time-consuming. In semantic segmentation particularly, it is attractive to train models on annotated images from a simulated (source) domain and deploy them on real (target) domains. In this work, we present a novel framework for unsupervised domain adaptation based on the notion of target-domain consistency training. Intuitively, our work is based on the insight that in order to perform well on the target domain, a model's output should be consistent with respect to small perturbations of inputs in the target domain. Specifically, we introduce a new loss term to enforce pixelwise consistency between the model's predictions on a target image and perturbed version of the same image. In comparison to popular adversarial adaptation methods, our approach is simpler, easier to implement, and more memory-efficient during training. Experiments and ablation studies demonstrate that our simple approach achieves remarkably strong results on two challenging synthetic-to-real benchmarks, GTA5-to-Cityscapes and SYNTHIA-to-Cityscapes.
****
Deep RGB-D Saliency Detection With Depth-Sensitive Attention and Automatic Multi-Modal Fusion
Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1407-1417
RGB-D salient object detection (SOD) is usually formulated as a problem of classification or regression over two modalities, i.e., RGB and depth. Hence, effective RGB-D feature modeling and multi-modal feature fusion both play a vital role

in RGB-D SOD. In this paper, we propose a depth-sensitive RGB feature modeling scheme using the depth-wise geometric prior of salient objects. In principle, the feature modeling scheme is carried out in a depth-sensitive attention module, which leads to the RGB feature enhancement as well as the background distraction reduction by capturing the depth geometry prior. Moreover, to perform effective multi-modal feature fusion, we further present an automatic architecture search approach for RGB-D SOD, which does well in finding out a feasible architecture from our specially designed multi-modal multi-scale search space. Extensive experiments on seven standard benchmarks demonstrate the effectiveness of the proposed approach against the state-of-the-art.
****
Exploring Sparsity in Image Super-Resolution for Efficient Inference
Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4917-4926
Current CNN-based super-resolution (SR) methods process all locations equally with computational resources being uniformly assigned in space. However, since missing details in low-resolution (LR) images mainly exist in regions of edges and textures, less computational resources are required for those flat regions. Therefore, existing CNN-based methods involve redundant computation in flat regions, which increases their computational cost and limits their applications on mobile devices. In this paper, we explore the sparsity in image SR to improve inference efficiency of SR networks. Specifically, we develop a Sparse Mask SR (SMSR) network to learn sparse masks to prune redundant computation. Within our SMSR, spatial masks learn to identify "important" regions while channel masks learn to mark redundant channels in those "unimportant" regions. Consequently, redundant computation can be accurately localized and skipped while maintaining comparable performance. It is demonstrated that our SMSR achieves state-of-the-art performance with 41%/33%/27% FLOPs being reduced for x2/3/4 SR. Code is available at: https://github.com/LongguangWang/SMSR.
****
Positive Sample Propagation Along the Audio-Visual Event Line
Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8436-8444
Visual and audio signals often coexist in natural environments, forming audio-visual events (AVEs). Given a video, we aim to localize video segments containing an AVE and identify its category. In order to learn discriminative features for a classifier, it is pivotal to identify the helpful (or positive) audio-visual segment pairs while filtering out the irrelevant ones, regardless whether they are synchronized or not. To this end, we propose a new positive sample propagation (PSP) module to discover and exploit the closely related audio-visual pairs by evaluating the relationship within every possible pair. It can be done by constructing an all-pair similarity map between each audio and visual segment, and only aggregating the features from the pairs with high similarity scores. To encourage the network to extract high correlated features for positive samples, a new audio-visual pair similarity loss is proposed. We also propose a new weighting branch to better exploit the temporal correlations in weakly supervised setting. We perform extensive experiments on the public AVE dataset and achieve new state-of-the-art accuracy in both fully and weakly supervised settings, thus verifying the effectiveness of our method.
****
Understanding the Behaviour of Contrastive Loss
Feng Wang, Huaping Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2495-2504
Unsupervised contrastive learning has achieved outstanding success, while the mechanism of contrastive loss has been less studied. In this paper, we concentrate on the understanding of the behaviours of unsupervised contrastive loss. We will show that the contrastive loss is a hardness-aware loss function, and the temperature t controls the strength of penalties on hard negative samples. The previ

ous study has shown that uniformity is a key property of contrastive learning. We build relations between the uniformity and the temperature t. We will show that uniformity helps the contrastive learning to learn separable features, however excessive pursuit to the uniformity makes the contrastive loss not tolerant to semantically similar samples, which may break the underlying semantic structure and be harmful to the formation of features useful for downstream tasks. This is caused by the inherent defect of the instance discrimination objective. Specifically, instance discrimination objective tries to push all different instances apart, ignoring the underlying relations between samples. Pushing semantically consistent samples apart has no positive effect for acquiring a prior informative to general downstream tasks. A well-designed contrastive loss should have some extents of tolerance to the closeness of semantically similar samples. Therefore, we find that the contrastive loss meets a uniformity-tolerance dilemma, and a good choice of temperature can compromise these two properties properly to both learn separable features and tolerant to semantically similar samples, improving the feature qualities and the downstream performances.
****
Variational Prototype Learning for Deep Face Recognition
Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11906-11915
Deep face recognition has achieved remarkable improvements due to the introduction of margin-based softmax loss, in which the prototype stored in the last linear layer represents the center of each class. In these methods, training samples are enforced to be close to positive prototypes and far apart from negative prototypes by a clear margin. However, we argue that prototype learning only employs sample-to-prototype comparisons without considering sample-to-sample comparisons during training and the low loss value gives us an illusion of perfect feature embedding, impeding the further exploration of SGD. To this end, we propose Variational Prototype Learning (VPL), which represents every class as a distribution instead of a point in the latent space. By identifying the slow feature drift phenomenon, we directly inject memorized features into prototypes to approximate variational prototype sampling. The proposed VPL can simulate sample-to-sample comparisons within the classification framework, encouraging the SGD solver to be more exploratory, while boosting performance. Moreover, VPL is conceptually simple, easy to implement, computationally efficient and memory saving. We present extensive experimental results on popular benchmarks, which demonstrate the superiority of the proposed VPL method over the state-of-the-art competitors.
****
StylePeople: A Generative Model of Fullbody Human Avatars
Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, Victor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5151-5160
We propose a new type of full-body human avatars, which combines parametric mesh-based body model with a neural texture. We show that with the help of neural textures, such avatars can successfully model clothing and hair, which usually poses a problem for mesh-based approaches. We also show how these avatars can be created from multiple frames of a video using backpropagation. We then propose a generative model for such avatars that can be trained from datasets of images and videos of people. The generative model allows us to sample random avatars as well as to create dressed avatars of people from one or few images.
****
Optimal Quantization Using Scaled Codebook
Yerlan Idelbayev, Pavlo Molchanov, Maying Shen, Hongxu Yin, Miguel A. Carreira-Perpinan, Jose M. Alvarez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12095-12104
We study the problem of quantizing N sorted, scalar datapoints with a fixed codebook containing K entries that are allowed to be rescaled. The problem is defined as finding the optimal scaling factor \alpha and the datapoint assignments into the \alpha-scaled codebook to minimize the squared error between original and

quantized points. Previously, the globally optimal algorithms for this problem w
ere derived only for certain codebooks (binary and ternary) or under the assumpt
ion of certain distributions (Gaussian, Laplacian). By studying the properties o
f the optimal quantizer, we derive an \calO(NK \log K) algorithm that is guarant
eed to find the optimal quantization parameters for any fixed codebook regardles
s of data distribution. We apply our algorithm to synthetic and real-world neura
l network quantization problems and demonstrate the effectiveness of our approac
h.
****

RPN Prototype Alignment for Domain Adaptive Object Detector
Yixin Zhang, Zilei Wang, Yushi Mao; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2021, pp. 12425-12434
Recent years have witnessed great progress in object detection. However, due to
the domain shift problem, applying the knowledge of an object detector learned f
rom one specific domain to another one often suffers severe performance degradat
ion. Most existing methods adopt feature alignment either on the backbone networ
k or instance classifier to increase the transferability of object detector. Dif
ferent from existing methods, we propose to perform feature alignment of foregro
und and background in the RPN stage such that the foreground and background RPN
proposals in target domain can be effectively separated. Specifically, we first
construct one set of learnable RPN prototypes, and then enforce the RPN features
to align with the prototypes for both source and target domains. It essentially
cooperates the learning of RPN prototypes and features to align the source and
target RPN features. In this paradigm, the pseudo label of proposals in target d
omain need be first generated, and we propose a simple yet effective method suit
able for RPN feature alignment,i.e., using the filtered detection results to gui
de the pseudo label generation of RPN proposals by IoU. Furthermore, we adopt Gr
ad CAM to find the discriminative region within a proposal and use it to increas
e the discriminability of RPN features for alignment by spatially weighting. We
conduct extensive experiments on multiple cross-domain detection scenarios. The
results show the effectiveness of our proposed method against previous state-of-
the-art methods.
****

Dual Contradistinctive Generative Autoencoder
Gaurav Parmar, Dacheng Li, Kwonjoon Lee, Zhuowen Tu; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 823-832
We present a new generative autoencoder model with dual contradistinctive losses
to improve generative autoencoder that performs simultaneous inference (reconst
ruction) and synthesis (sampling). Our model, named dual contradistinctive gener
ative autoencoder (DC-VAE), integrates an instance-level discriminative loss (ma
intaining the instancelevel fidelity for the reconstruction / synthesis) with a
set-level adversarial loss (encouraging the set-level fidelity for the reconstru
ction/synthesis), both being contradistinctive. Extensive experimental results b
y DC-VAE across different resolutions including 32x32, 64x64, 128x128, and 512x5
12 are reported. The two contradistinctive losses in VAE work harmoniously in DC
-VAE leading to a significant qualitative and quantitative performance enhanceme
nt over the baseline VAEs without architectural changes. State-of-the-art or com
petitive results among generative autoencoders for image reconstruction, image s
ynthesis, image interpolation, and representation learning are observed. DC-VAE
is a general-purpose VAE model, applicable to a wide variety of downstream tasks
in computer vision and machine learning.
****

Binary TTC: A Temporal Geofence for Autonomous Navigation
Abhishek Badki, Orazio Gallo, Jan Kautz, Pradeep Sen; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12946-
12955
Time-to-contact (TTC), the time for an object to collide with the observer's pla
ne, is a powerful tool for path planning: it is potentially more informative tha
n the depth, velocity, and acceleration of objects in the scene---even for human
s. TTC presents several advantages, including requiring only a monocular, uncali

brated camera. However, regressing TTC for each pixel is not straightforward, an
d most existing methods make over-simplifying assumptions about the scene. We ad
dress this challenge by estimating TTC via a series of simpler, binary classific
ations. We predict with low latency whether the observer will collide with an ob
stacle within a certain time, which is often more critical than knowing exact, p
er-pixel TTC. For such scenarios, our method offers a temporal geofence in 6.4 m
s---over 25x faster than existing methods. Our approach can also estimate per-pi
xel TTC with arbitrarily fine quantization (including continuous values), when t
he computational budget allows for it. To the best of our knowledge, our method
is the first to offer TTC information (binary or coarsely quantized) at sufficie
ntly high frame-rates for practical use.
****

Semantic-Aware Video Text Detection
Wei Feng, Fei Yin, Xu-Yao Zhang, Cheng-Lin Liu; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1695-1705
Most existing video text detection methods track texts with appearance features,
 which are easily influenced by the change of perspective and illumination. Comp
ared with appearance features, semantic features are more robust cues for matchi
ng text instances. In this paper, we propose an end-to-end trainable video text
detector that tracks texts based on semantic features. First, we introduce a new
 character center segmentation branch to extract semantic features, which encode
 the category and position of characters. Then we propose a novel appearance-sem
antic-geometry descriptor to track text instances, in which semantic features ca
n improve the robustness against appearance changes. To overcome the lack of cha
racter-level annotations, we propose a novel weakly-supervised character center
detection module, which only uses word-level annotated real images to generate c
haracter-level labels. The proposed method achieves state-of-the-art performance
 on three video text benchmarks ICDAR 2013 Video, Minetto and RT-1K, and two Chi
nese scene text benchmarks CASIA10K and MSRA-TD500.
****

Real-Time High-Resolution Background Matting
Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L. Curless, Steven M.
Seitz, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2021, pp. 8762-8771
We introduce a real-time, high-resolution background replacement technique which
 operates at 30fps in 4K resolution, and 60fps for HD on a modern GPU. Our techn
ique is based on background matting, where an additional frame of the background
 is captured and used to inform the alpha matte and the foreground layer. The ma
in challenge is to compute a high-quality alpha matte, preserving strand-level h
air details, while processing high-resolution images in real-time. To achieve th
is goal, we employ two neural networks; the base network computes a low-resoluti
on result which is refined by a second network operating at high-resolution on s
elective patches. We introduce two large-scale video and image matting datasets:
 VideoMatte240K and PhotoMatte13K/85. Our approach yields higher quality results
 compared to the previous state-of-the-art in background matting, while simultan
eously yielding a dramatic boost in both speed and resolution.
****

Interpretable Social Anchors for Human Trajectory Forecasting in Crowds
Parth Kothari, Brian Sifringer, Alexandre Alahi; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15556-15566
Human trajectory forecasting in crowds, at its core, is a sequence prediction pr
oblem with specific challenges of capturing inter-sequence dependencies (social
interactions) and consequently predicting socially-compliant multimodal distribu
tions. In recent years, neural network-based methods have been shown to outperfo
rm hand-crafted methods on distance-based metrics. However, these data-driven me
thods still suffer from one crucial limitation: lack of interpretability. To ove
rcome this limitation, we leverage the power of discrete choice models to learn
interpretable rule-based intents, and subsequently utilise the expressibility of
 neural networks to model scene-specific residual. Extensive experimentation on
the interaction-centric benchmark TrajNet++ demonstrates the effectiveness of ou

r proposed architecture to explain its predictions without compromising the accu racy.
****
Trajectory Prediction With Latent Belief Energy-Based Model
Bo Pang, Tianyang Zhao, Xu Xie, Ying Nian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11814-11824
Human trajectory prediction is critical for autonomous platforms like self-driving cars or social robots. We present a latent belief energy-based model (LB-EBM) for diverse human trajectory forecast. LB-EBM is a probabilistic model with cost function defined in the latent space to account for the movement history and social context. The low-dimensionality of the latent space and the high expressivity of the EBM make it easy for the model to capture the multimodality of pedestrian trajectory distributions. LB-EBM is learned from expert demonstrations (i.e., human trajectories) projected into the latent space. Sampling from or optimizing the learned LB-EBM yields a belief vector which is used to make a path plan, which then in turn helps to predict a long-range trajectory. The effectiveness of LB-EBM and the two-step approach are supported by strong empirical results. Our model is able to make accurate, multi-modal, and social compliant trajectory predictions and improves over prior state-of-the-arts performance on the Stanford Drone trajectory prediction benchmark by 10.9% and on the ETH-UCY benchmark by 27.6%.
****
Metadata Normalization
Mandy Lu, Qingyu Zhao, Jiequan Zhang, Kilian M. Pohl, Li Fei-Fei, Juan Carlos Niebles, Ehsan Adeli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10917-10927
Batch Normalization (BN) and its variants have delivered tremendous success in combating the covariate shift induced by the training step of deep learning methods. While these techniques normalize the feature distribution by standardizing with batch statistics, they do not correct the influence on features from extraneous variables or multiple distributions. Such extra variables, referred to as metadata here, may create bias or confounding effects (e.g., race when classifying gender from face images). We introduce the Metadata Normalization (MDN) layer, a new batch-level operation which can be used end-to-end within the training framework, to correct the influence of metadata on the feature distribution. MDN adopts a regression analysis technique traditionally used for preprocessing to remove (regress out) the metadata effects on model features during training. We utilize a metric based on distance correlation to quantify the distribution bias from the metadata and demonstrate that our method successfully removes metadata effects on four diverse settings: one synthetic, one 2D image, one video, and one 3D medical image dataset.
****
Multi-Objective Interpolation Training for Robustness To Label Noise
Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor, Kevin McGuinness; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6606-6615
Deep neural networks trained with standard cross-entropy loss memorize noisy labels, which degrades their performance. Most research to mitigate this memorization proposes new robust classification loss functions. Conversely, we propose a Multi-Objective Interpolation Training (MOIT) approach that jointly exploits contrastive learning and classification to mutually help each other and boost performance against label noise. We show that standard supervised contrastive learning degrades in the presence of label noise and propose an interpolation training strategy to mitigate this behavior. We further propose a novel label noise detection method that exploits the robust feature representations learned via contrastive learning to estimate per-sample soft-labels whose disagreements with the original labels accurately identify noisy samples. This detection allows treating noisy samples as unlabeled and training a classifier in a semi-supervised manner to prevent noise memorization and improve representation learning. We further propose MOIT+, a refinement of MOIT by fine-tuning on detected clean samples. Hype

rparameter and ablation studies verify the key components of our method. Experim
ents on synthetic and real-world noise benchmarks demonstrate that MOIT/MOIT+ ac
hieves state-of-the-art results. Code is available at https://git.io/JI40X.
****
PhySG: Inverse Rendering With Spherical Gaussians for Physics-Based Material Edi
ting and Relighting
Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, Noah Snavely; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
 pp. 5453-5462
We present an end-to-end inverse rendering pipeline that includes a fully differ
entiable renderer, and can reconstruct geometry, materials, and illumination fro
m scratch from a set of images. Our rendering framework represents specular BRDF
s and environmental illumination using mixtures of spherical Gaussians, and repr
esents geometry as a signed distance function parameterized as a Multi-Layer Per
ceptron. The use of spherical Gaussians allows us to efficiently solve for appro
ximate light transport, and our method works on scenes with challenging non-Lamb
ertian reflectance captured under natural, static illumination. We demonstrate,
with both synthetic and real data, that our reconstruction not only can render n
ovel viewpoints, but also enables physics-based appearance editing of materials
and illumination.
****
Predator: Registration of 3D Point Clouds With Low Overlap
Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, Konrad Schindler;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2021, pp. 4267-4276
We introduce PREDATOR, a model for pairwise pointcloud registration with deep at
tention to the overlap region. Different from previous work, our model is specif
ically designed to handle (also) point-cloud pairs with low overlap. Its key nov
elty is an overlap-attention block for early information exchange between the la
tent encodings of the two point clouds. In this way the subsequent decoding of t
he latent representations into per-point features is conditioned on the respecti
ve other point cloud, and thus can predict which points are not only salient, bu
t also lie in the overlap region between the two point clouds. The ability to fo
cus on points that are relevant for matching greatly improves performance: PREDA
TOR raises the rate of successful registrations by more than 20% in the low-over
lap scenario, and also sets a new state of the art for the 3DMatch benchmark wit
h 89% registration recall.
****
Hierarchical Motion Understanding via Motion Programs
Sumith Kulal, Jiayuan Mao, Alex Aiken, Jiajun Wu; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6568-6576
Current approaches to video analysis of human motion focus on raw pixels or keyp
oints as the basic units of reasoning. We posit that adding higher-level motion
primitives, which can capture natural coarser units of motion such as backswing
or follow-through, can be used to improve downstream analysis tasks. This higher
 level of abstraction can also capture key features, such as loops of repeated p
rimitives, that are currently inaccessible at lower levels of representation. We
 therefore introduce Motion Programs, a neuro-symbolic, program-like representat
ion that expresses motions as a composition of high-level primitives. We also pr
esent a system for automatically inducing motion programs from videos of human m
otion and for leveraging motion programs in video synthesis. Experiments show th
at motion programs can accurately describe a diverse set of human motions and th
e inferred programs contain semantically meaningful motion primitives, such as a
rm swings and jumping jacks. Our representation also benefits downstream tasks s
uch as video interpolation and video prediction and outperforms off-the-shelf mo
dels. We further demonstrate how these programs can detect diverse kinds of repe
titive motion and facilitate interactive video editing.
****
Neural Side-by-Side: Predicting Human Preferences for No-Reference Super-Resolut
ion Evaluation

Valentin Khrulkov, Artem Babenko; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2021, pp. 4988-4997
Super-resolution based on deep convolutional networks is currently gaining much attention from both academia and industry. However, lack of proper evaluation me asures makes it difficult to compare approaches, hampering progress in the field . Traditional measures, such as PSNR or SSIM, are known to poorly correlate with the human perception of image quality. Therefore, in existing works common prac tice is also to report Mean-Opinion-Score (MOS) -- the results of human evaluati on of super-resolved images. Unfortunately, the MOS values from different papers are not directly comparable, due to the varying number of raters, their subject ivity, etc. By this paper, we introduce Neural Side-By-Side -- a new measure tha t allows super-resolution models to be compared automatically, effectively appro ximating human preferences. Namely, we collect a large dataset of aligned image pairs, which were produced by different super-resolution models. Then each pair is annotated by several raters, who were instructed to choose a more visually ap pealing image. Given the dataset and the labels, we trained a CNN model that obt ains a pair of images and for each image predicts a probability of being more pr eferable than its counterpart. In this work, we show that Neural Side-By-Side ge neralizes across both new models and new data. Hence, it can serve as a natural approximation of human preferences, which can be used to compare models or tune hyperparameters without raters' assistance. We open-source the dataset and the p retrained model and expect that it will become a handy tool for researchers and practitioners.
****
Coordinate Attention for Efficient Mobile Network Design
Qibin Hou, Daquan Zhou, Jiashi Feng; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2021, pp. 13713-13722
Recent studies on mobile network design have demonstrated the remarkable effecti veness of channel attention (e.g., the Squeeze-and-Excitation attention) for lif ting model performance, but they generally neglect the positional information, w hich is important for generating spatially selective attention maps. In this pap er, we propose a novel attention mechanism for mobile networks by embedding posi tional information into channel attention, which we call "coordinate attention". Unlike channel attention that transforms a feature tensor to a single feature v ector via 2D global pooling, the coordinate attention factorizes channel attenti on into two 1D feature encoding processes that aggregate features along the two spatial directions, respectively. In this way, long-range dependencies can be ca ptured along one spatial direction and meanwhile precise positional information can be preserved along the other spatial direction. The resulting feature maps a re then encoded separately into a pair of direction-aware and position-sensitive attention maps that can be complementarily applied to the input feature map to augment the representations of the objects of interest. Our coordinate attention is simple and can be flexibly plugged into classic mobile networks, such as Mob ileNetV2, MobileNeXt, and EfficientNet with nearly no computational overhead. Ex tensive experiments demonstrate that our coordinate attention is not only benefi cial to ImageNet classification but more interestingly, behaves better in down-s tream tasks, such as object detection and semantic segmentation. Code is availab le at https://github.com/Andrew-Qibin/CoordAttention.
****
Stylized Neural Painting
Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, Zhenwei Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp . 15689-15698
This paper proposes an image-to-painting translation method that generates vivid and realistic painting artworks with controllable styles. Different from previo us image-to-image translation methods that formulate the translation as pixel-wi se prediction, we deal with such an artistic creation process in a vectorized en vironment and produce a sequence of physically meaningful stroke parameters that can be further used for rendering. Since a typical vector render is not differe ntiable, we design a novel neural renderer which imitates the behavior of the ve

ctor renderer and then frame the stroke prediction as a parameter searching process that maximizes the similarity between the input and the rendering output. We explored the zero-gradient problem on parameter searching and propose to solve this problem from an optimal transportation perspective. We also show that previous neural renderers have a parameter coupling problem and we re-design the rendering network with a rasterization network and a shading network that better handles the disentanglement of shape and color. Experiments show that the paintings generated by our method have a high degree of fidelity in both global appearance and local textures. Our method can be also jointly optimized with neural style transfer that further transfers visual style from other images. Our code and animated results are available at https://jiupinjia.github.io/neuralpainter/.
****
Image Change Captioning by Learning From an Auxiliary Task
Mehrdad Hosseinzadeh, Yang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2725-2734
We tackle the challenging task of image change captioning. The goal is to describe the subtle difference between two very similar images by generating a sentence caption. While the recent methods mainly focus on proposing new model architectures for this problem, we instead focus on an alternative training scheme. Inspired by the success of multi-task learning, we formulate a training scheme that uses an auxiliary task to improve the training of the change captioning network. We argue that the task of composed query image retrieval is a natural choice as the auxiliary task. Given two almost similar images as the input, the primary network generates a caption describing the fine change between those two images. Next, the auxiliary network is provided with the generated caption and one of those two images. It then tries to pick the second image among a set of candidates. This forces the primary network to generate detailed and precise captions via having an extra supervision loss by the auxiliary network. Furthermore, we propose a new scheme for selecting a negative set of candidates for the retrieval task that can effectively improve the performance. We show that the proposed training strategy performs well on the task of change captioning on benchmark datasets.
****
Learning to Generalize Unseen Domains via Memory-based Multi-Source Meta-Learning for Person Re-Identification
Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6277-6286
Recent advances in person re-identification (ReID) obtain impressive accuracy in the supervised and unsupervised learning settings. However, most of the existing methods need to train a new model for a new domain by accessing data. Due to public privacy, the new domain data are not always accessible, leading to a limited applicability of these methods. In this paper, we study the problem of multi-source domain generalization in ReID, which aims to learn a model that can perform well on unseen domains with only several labeled source domains. To address this problem, we propose the Memory-based Multi-Source Meta-Learning (M^3L) framework to train a generalizable model for unseen domains. Specifically, a meta-learning strategy is introduced to simulate the train-test process of domain generalization for learning more generalizable models. To overcome the unstable meta-optimization caused by the parametric classifier, we propose a memory-based identification loss that is non-parametric and harmonizes with meta-learning. We also present a meta batch normalization layer (MetaBN) to diversify meta-test features, further establishing the advantage of meta-learning. Experiments demonstrate that our M^3L can effectively enhance the generalization ability of the model for unseen domains and can outperform the state-of-the-art methods on four large-scale ReID datasets.
****
Discriminative Appearance Modeling With Multi-Track Pooling for Real-Time Multi-Object Tracking
Chanho Kim, Li Fuxin, Mazen Alotaibi, James M. Rehg; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9553-9562
In multi-object tracking, the tracker maintains in its memory the appearance and motion information for each object in the scene. This memory is utilized for finding matches between tracks and detections, and is updated based on the matching. Many approaches model each target in isolation and lack the ability to use all the targets in the scene to jointly update the memory. This can be problematic when there are similarly looking objects in the scene. In this paper, we solve the problem of simultaneously considering all tracks during memory updating, with only a small spatial overhead, via a novel multi-track pooling module. We additionally propose a training strategy adapted to multi-track pooling which generates hard tracking episodes online. We show that the combination of these innovations results in a strong discriminative appearance model under the bilinear LSTM tracking framework, enabling the use of greedy data association to achieve online tracking performance. Our experiments demonstrate real-time, state-of-the-art online tracking performance on public multi-object tracking (MOT) datasets.
****
LASR: Learning Articulated Shape Reconstruction From a Monocular Video
Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T. Freeman, Ce Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15980-15989
Remarkable progress has been made in 3D reconstruction of rigid structures from a video or a collection of images. However, it is still challenging to reconstruct nonrigid structures from RGB inputs, due to the under-constrained nature of this problem. While template-based approaches, such as parametric shape models, have achieved great success in terms of modeling the "closed world" of known object categories, their ability to handle the "open-world" of novel object categories and outlier shapes is still limited. In this work, we introduce a template-free approach for 3D shape learning from a single video. It adopts an analysis-by-synthesis strategy that forward-renders object silhouette, optical flow, and pixels intensities to compare against video observations, which generates gradients signals to adjust the camera, shape and motion parameters. Without relying on a category-specific shape template, our method faithfully reconstructs nonrigid 3D structures from videos of human, animals, and objects of unknown classes in the wild.
****
FVC: A New Framework Towards Deep Video Compression in Feature Space
Zhihao Hu, Guo Lu, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1502-1511
Learning based video compression attracts increasing attention in the past few years. The previous hybrid coding approaches rely on pixel space operations to reduce spatial and temporal redundancy, which may suffer from inaccurate motion estimation or less effective motion compensation. In this work, we propose a feature-space video coding network (FVC) by performing all major operations (i.e., motion estimation, motion compression, motion compensation and residual compression) in the feature space. Specifically, in the proposed deformable compensation module, we first apply motion estimation in the feature space to produce motion information (i.e., the offset maps), which will be compressed by using the auto-encoder style network. Then we perform motion compensation by using deformable convolution and generate the predicted feature. After that, we compress the residual feature between the feature from the current frame and the predicted feature from our deformable compensation module. For better frame reconstruction, the reference features from multiple previous reconstructed frames are also fused by using the non-local attention mechanism in the multi-frame feature fusion module. Comprehensive experimental results demonstrate that the proposed framework achieves the state-of-the-art performance on four benchmark datasets including HEVC, UVG, VTL and MCL-JCV.
****
Exponential Moving Average Normalization for Self-Supervised and Semi-Supervised Learning

Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 194-203
We present a plug-in replacement for batch normalization (BN) called exponential moving average normalization (EMAN), which improves the performance of existing student-teacher based self- and semi-supervised learning techniques. Unlike the standard BN, where the statistics are computed within each batch, EMAN, used in the teacher, updates its statistics by exponential moving average from the BN statistics of the student. This design reduces the intrinsic cross-sample dependency of BN and enhances the generalization of the teacher. EMAN improves strong baselines for self-supervised learning by 4-6/1-2 points and semi-supervised learning by about 7/2 points, when 1%/10% supervised labels are available on ImageNet. These improvements are consistent across methods, network architectures, training duration, and datasets, demonstrating the general effectiveness of this technique. The code will be made available online.
****
Confluent Vessel Trees With Accurate Bifurcations
Zhongwen Zhang, Dmitrii Marin, Maria Drangova, Yuri Boykov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9573-9582
We are interested in unsupervised reconstruction of complex near-capillary vasculature with thousands of bifurcations where supervision and learning are infeasible. Unsupervised methods can use many structural constraints, e.g. topology, geometry, physics. Common techniques use variants of MST on geodesic "tubular graphs" minimizing symmetric pairwise costs, i.e. distances. We show limitations of such standard undirected tubular graphs producing typical errors at bifurcations where flow "directedness" is critical. We introduce a new general concept of "confluence" for continuous oriented curves forming vessel trees and show how to enforce it on discrete tubular graphs. While confluence is a high-order property, we present an efficient practical algorithm for reconstructing confluent vessel trees using minimum arborescence on a directed graph enforcing confluence via simple flow-extrapolating arc construction. Empirical tests on large near-capillary sub-voxel vasculature volumes demonstrate significantly improved reconstruction accuracy at bifurcations. Our code has also been made publicly available.
****
Intentonomy: A Dataset and Study Towards Human Intent Understanding
Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12986-12996
An image is worth a thousand words, conveying information that goes beyond the physical visual content therein. In this paper, we study the intent behind social media images with an aim to analyze how visual information can help the recognition of human intent. Towards this goal, we introduce an intent dataset, Intentonomy, comprising 14K images covering a wide range of everyday scenes. These images are manually annotated with 28 intent categories that are derived from a social psychology taxonomy. We then systematically study whether, and to what extent, commonly used visual information, i.e., object and context, contribute to human motive understanding. Based on our findings, we conduct further study to quantify the effect of attending to object and context classes as well as textual information in the form of hashtags when training an intent classifier. Our results quantitatively and qualitatively shed light on how visual and textual information can produce observable effects when predicting intent.
****
End-to-End Rotation Averaging With Multi-Source Propagation
Luwei Yang, Heng Li, Jamal Ahmed Rahim, Zhaopeng Cui, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11774-11783
This paper presents an end-to-end neural network for multiple rotation averaging in SfM. Due to the manifold constraint of rotations, conventional methods usually take two separate steps involving spanning tree based initialization and iter

ative nonlinear optimization respectively. These methods can suffer from bad initializations due to the noisy spanning tree or outliers in input relative rotations. To handle these problems, we propose to integrate initialization and optimization together in an unified graph neural network via a novel differentiable multi-source propagation module. Specifically, our network utilizes image context and geometric cues in feature correspondences to reduce the impact of outliers. Furthermore, unlike the methods that utilize the spanning tree to initialize orientations according to a single reference node in a top-down manner, our network initializes orientations according to multiple sources while utilizing information from all neighbors in a differentiable way.More importantly, our end-to-end formulation also enables iterative re-weighting of input relative orientations at test time to improve the accuracy of the final estimation by minimizing the impact of outliers. We demonstrate the effectiveness of our method on two real-world datasets, achieving state-of-the-art performance.
****
Controllable Image Restoration for Under-Display Camera in Smartphones
Kinam Kwon, Eunhee Kang, Sangwon Lee, Su-Jin Lee, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2073-2082
Under-display camera (UDC) technology is essential for full-screen display in smartphones and is achieved by removing the concept of drilling holes on display. However, this causes inevitable image degradation in the form of spatially variant blur and noise because of the opaque display in front of the camera. To address spatially variant blur and noise in UDC images, we propose a novel controllable image restoration algorithm utilizing pixel-wise UDC-specific kernel representation and a noise estimator. The kernel representation is derived from an elaborate optical model that reflects the effect of both normal and oblique light incidence. Also, noise-adaptive learning is introduced to control noise levels, which can be utilized to provide optimal results depending on the user preferences. The experiments showed that the proposed method achieved superior quantitative performance as well as higher perceptual quality on both a real-world dataset and a monitor-based aligned dataset compared to conventional image restoration algorithms.
****
Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification
Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1522-1531
The Information Bottleneck (IB) provides an information theoretic principle for representation learning, by retaining all information relevant for predicting label while minimizing the redundancy. Though IB principle has been applied to a wide range of applications, its optimization remains a challenging problem which heavily relies on the accurate estimation of mutual information. In this paper, we present a new strategy, Variational Self-Distillation (VSD), which provides a scalable, flexible and analytic solution to essentially fitting the mutual information but without explicitly estimating it. Under rigorously theoretical guarantee, VSD enables the IB to grasp the intrinsic correlation between representation and label for supervised training. Furthermore, by extending VSD to multi-view learning, we introduce two other strategies, Variational Cross-Distillation (VCD) and Variational Mutual Learning (VML), which significantly improve the robustness of representation to view-changes by eliminating view-specific and task-irrelevant information. To verify our theoretically grounded strategies, we apply our approaches to cross-modal person Re-ID, and conduct extensive experiments, where the superior performance against state-of-the-art methods are demonstrated. Our intriguing findings highlight the need to rethink the way to estimate mutual information.
****
Context-Aware Biaffine Localizing Network for Temporal Sentence Grounding
Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu,

Yulai Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11235-11244
This paper addresses the problem of temporal sentence grounding (TSG), which aims to identify the temporal boundary of a specific segment from an untrimmed video by a sentence query. Previous works either compare pre-defined candidate segments with the query and select the best one by ranking, or directly regress the boundary timestamps of the target segment. In this paper, we propose a novel localization framework that scores all pairs of start and end indices within the video simultaneously with a biaffine mechanism. In particular, we present a Context-aware Biaffine Localizing Network (CBLN) which incorporates both local and global contexts into features of each start/end position for biaffine-based localization. The local contexts from the adjacent frames help distinguish the visually similar appearance, and the global contexts from the entire video contribute to reasoning the temporal relation. Besides, we also develop a multi-modal self-attention module to provide fine-grained query-guided video representation for this biaffine strategy. Extensive experiments show that our CBLN significantly outperforms state-of-the-arts on three public datasets (ActivityNet Captions, TACoS, and Charades-STA), demonstrating the effectiveness of the proposed localization framework.
****
NewtonianVAE: Proportional Control and Goal Identification From Pixels via Physical Latent Spaces
Miguel Jaques, Michael Burke, Timothy M. Hospedales; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4454-4463
Learning low-dimensional latent state space dynamics models has proven powerful for enabling vision-based planning and learning for control. We introduce a latent dynamics learning framework that is uniquely designed to induce proportional controlability in the latent space, thus enabling the use of simple and well-known PID controllers. We show that our learned dynamics model enables proportional control from pixels, dramatically simplifies and accelerates behavioural cloning of vision-based controllers, and provides interpretable goal discovery when applied to imitation learning of switching controllers from demonstration. Notably, such proportional controlability also allows for robust path following from visual demonstrations using Dynamic Movement Primitives in the learned latent space.
****
Auto-Exposure Fusion for Single-Image Shadow Removal
Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10571-10580
Shadow removal is still a challenging task due to its inherent background-dependent and spatial-variant properties, leading to unknown and diverse shadow patterns. Even powerful deep neural networks could hardly recover traceless shadow-removed background. This paper proposes a new solution for this task by formulating it as an exposure fusion problem to address the challenges. Intuitively, we first estimate multiple over-exposure images w.r.t. the input image to let the shadow regions in these images have the same color with shadow-free areas in the input image. Then, we fuse the original input with the over-exposure images to generate the final shadow-free counterpart. Nevertheless, the spatial-variant property of the shadow requires the fusion to be sufficiently `smart', that is, it should automatically select proper over-exposure pixels from different images to make the final output natural. To address this challenge, we propose the shadow-aware FusionNet that takes the shadow image as input to generate fusion weight maps across all the over-exposure images. Moreover, we propose the boundary-aware RefineNet to eliminate the remaining shadow trace further. We conduct extensive experiments on the ISTD, ISTD+, and SRD datasets to validate our method's effectiveness and show better performance in shadow regions and comparable performance in non-shadow regions over the state-of-the-art methods. We release the code in https://github.com/tsingqguo/exposure-fusion-shadow-removal.

****

Anticipating Human Actions by Correlating Past With the Future With Jaccard Similarity Measures

Basura Fernando, Samitha Herath; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13224-13233

We propose a framework for early action recognition and anticipation by correlating past features with the future using three novel similarity measures called Jaccard vector similarity, Jaccard cross-correlation and Jaccard Frobenius inner product over covariances. Using these combinations of novel losses and using our framework, we obtain state-of-the-art results for early action recognition in UCF101 and JHMDB datasets by obtaining 91.7 % and 83.5 % accuracy respectively for an observation percentage of 20. Similarly, we obtain state-of-the-art results for Epic-Kitchen55 and Breakfast datasets for action anticipation by obtaining 20.35 and 41.8 top-1 accuracy respectively.

****

LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces From Video Using Pose and Lighting Normalization

Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, Chris Bregler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2755-2764

In this paper, we present a video-based learning framework for animating personalized 3D talking faces from audio. We introduce two training-time data normalizations that significantly improve data sample efficiency. First, we isolate and represent faces in a normalized space that decouples 3D geometry, head pose, and texture. This decomposes the prediction problem into regressions over the 3D face shape and the corresponding 2D texture atlas. Second, we leverage facial symmetry and approximate albedo constancy of skin to isolate and remove spatiotemporal lighting variations. Together, these normalizations allow simple networks to generate high fidelity lip-sync videos under novel ambient illumination while training with just a single video (of usually < 5 minutes). Further, to stabilize temporal dynamics, we introduce an auto-regressive approach that conditions the model on its previous visual state. Human ratings and objective metrics demonstrate that our method outperforms contemporary state-of-the-art audio-driven video reenactment benchmarks in terms of realism, lip-sync and visual quality scores. We illustrate several applications enabled by our framework.

****

Simpler Certified Radius Maximization by Propagating Covariances

Xingjian Zhen, Rudrasis Chakraborty, Vikas Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7292-7301

One strategy for adversarially training a robust model is to maximize its certified radius -- the neighborhood around a given training sample for which the model's prediction remains unchanged. The scheme typically involves analyzing a "smoothed" classifier where one estimates the prediction corresponding to Gaussian samples in the neighborhood of each sample in the mini-batch, accomplished in practice by Monte Carlo sampling. In this paper, we investigate the hypothesis that this sampling bottleneck can potentially be mitigated by identifying ways to directly propagate the covariance matrix of the smoothed distribution through the network. To this end, we find that other than certain adjustments to the network, propagating the covariances must also be accompanied by additional accounting that keeps track of how the distributional moments transform and interact at each stage in the network. We show how satisfying these criteria yields an algorithm for maximizing the certified radius on datasets including Cifar-10, ImageNet, and Places365 while offering runtime savings on networks with moderate depth, with a small compromise in overall accuracy. We describe the details of the key modifications that enable practical use. Via various experiments, we evaluate when our simplifications are sensible, and what the key benefits and limitations are.

****

A 3D GAN for Improved Large-Pose Facial Recognition

Richard T. Marriott, Sami Romdhani, Liming Chen; Proceedings of the IEEE/CVF Con

ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13445-13455
Facial recognition using deep convolutional neural networks relies on the availa bility of large datasets of face images. Many examples of identities are needed, and for each identity, a large variety of images are needed in order for the ne twork to learn robustness to intra-class variation. In practice, such datasets a re difficult to obtain, particularly those containing adequate variation of pose . Generative Adversarial Networks (GANs) provide a potential solution to this pr oblem due to their ability to generate realistic, synthetic images. However, rec ent studies have shown that current methods of disentangling pose from identity are inadequate. In this work we incorporate a 3D morphable model into the genera tor of a GAN in order to learn a nonlinear texture model from in-the-wild images . This allows generation of new, synthetic identities, and manipulation of pose, illumination and expression without compromising the identity. Our synthesised data is used to augment training of facial recognition networks with performance evaluated on the challenging CFP and CPLFW datasets.
****
Repopulating Street Scenes
Yifan Wang, Andrew Liu, Richard Tucker, Jiajun Wu, Brian L. Curless, Steven M. S eitz, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2021, pp. 5110-5119
We present a framework for automatically reconfiguring images of street scenes b y populating, depopulating, or repopulating them with objects such as pedestrian s or vehicles. Applications of this method include anonymizing images to enhance privacy, generating data augmentations for perception tasks like autonomous dri ving, and composing scenes to achieve a certain ambiance, such as empty streets in the early morning. At a technical level, our work has three primary contribut ions: (1) a method for clearing images of objects,(2) a method for estimating su n direction from a single image, and (3) a way to compose objects in scenes that respects scene geometry and illumination. Each component is learned from data w ith minimal ground truth annotations, by making creative use of large-numbers of short image bursts of street scenes. We demonstrate convincing results on a ran ge of street scenes and illustrate potential applications.
****
ARVo: Learning All-Range Volumetric Correspondence for Video Deblurring
Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suom inen, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7721-7731
Video deblurring models exploit consecutive frames to remove blurs from camera s hakes and object motions. In order to utilize neighboring sharp patches, typical methods rely mainly on homography or optical flows to spatially align neighbori ng blurry frames. However, such explicit approaches are less effective in the pr esence of fast motions with large pixel displacements. In this work, we propose a novel implicit method to learn spatial correspondence among blurry frames in t he feature space. To construct distant pixel correspondences, our model builds a correlation volume pyramid among all the pixel-pairs between neighboring frames . To enhance the features of the reference frame, we design a correlative aggreg ation module that maximizes the pixel-pair correlations with its neighbors based on the volume pyramid. Finally, we feed the aggregated features into a reconstr uction module to obtain the restored frame. We design a generative adversarial p aradigm to optimize the model progressively. Our proposed method is evaluated on the widely-adopted DVD dataset, along with a newly collected High-Frame-Rate (1 000 fps) Dataset for Video Deblurring (HFR-DVD). Quantitative and qualitative ex periments show that our model performs favorably on both datasets against previo us state-of-the-art methods, confirming the benefit of modeling all-range spatia l correspondence for video deblurring.
****
Unsupervised Object Detection With LIDAR Clues
Hao Tian, Yuntao Chen, Jifeng Dai, Zhaoxiang Zhang, Xizhou Zhu; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5962-5972

Despite the importance of unsupervised object detection, to the best of our know ledge, there is no previous work addressing this problem. One main issue, widely known to the community, is that object boundaries derived only from 2D image ap pearance are ambiguous and unreliable. To address this, we exploit LiDAR clues t o aid unsupervised object detection. By exploiting the 3D scene structure, the i ssue of localization can be considerably mitigated. We further identify another major issue, seldom noticed by the community, that the long-tailed and open-ende d (sub-)category distribution should be accommodated. In this paper, we present the first practical method for unsupervised object detection with the aid of LiD AR clues. In our approach, candidate object segments based on 3D point clouds ar e firstly generated. Then, an iterative segment labeling process is conducted to assign segment labels and to train a segment labeling network, which is based o n features from both 2D images and 3D point clouds. The labeling process is care fully designed so as to mitigate the issue of long-tailed and open-ended distrib ution. The final segment labels are set as pseudo annotations for object detecti on network training. Extensive experiments on the large-scale Waymo Open dataset suggest that the derived unsupervised object detection method achieves reasonab le accuracy compared with that of strong supervision within the LiDAR visible ra nge.
****
TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking
N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2021, pp. 15190-15200
We consider the task of 3D pose estimation and trackingof multiple people seen i n an arbitrary number of camerafeeds. We propose TesseTrack, a novel top-down ap proachthat simultaneously reasons about multiple individuals' 3Dbody joint recon structions and associations in space andtime in a single end-to-end learnable fr amework. At the core of our approach is a novel spatio-temporal formulation that operates in a common voxelized feature space aggregated from single- or multipl e-camera views. After a person detection step, a 4D CNN produces short-term pers on-specific representations which are then linked across time by a differentiabl e matcher. The linked descriptions are then merged and deconvolved into 3D poses . This joint spatio-temporal formulation contrasts with previous piece-wise stra tegies that treat 2D pose estimation, 2D-to-3D lifting, and 3D pose tracking as independent sub-problems that are error-prone when solved in isolation. Furtherm ore, unlike previous methods, TesseTrack is robust to changes in the number of c amera views and achieves very good results even if a single view is available at inference time. Quantitative evaluation of 3D pose reconstruction accuracy on s tandard benchmarks shows significant improvements over the state of the art. Eva luation of multi-person articulated 3D pose tracking in our novel evaluation fra mework demonstrates the superiority of TesseTrack over strong baselines.
****
HVPR: Hybrid Voxel-Point Representation for Single-Stage 3D Object Detection
Jongyoun Noh, Sanghoon Lee, Bumsub Ham; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14605-14614
We address the problem of 3D object detection, that is, estimating 3D object bou nding boxes from point clouds. 3D object detection methods exploit either voxel-based or point-based features to represent 3D objects in a scene. Voxel-based fe atures are efficient to extract, while they fail to preserve fine-grained 3D str uctures of objects. Point-based features, on the other hand, represent the 3D st ructures more accurately, but extracting these features is computationally expen sive. We introduce in this paper a novel single-stage 3D detection method having the merit of both voxel-based and point-based features. To this end, we propose a new convolutional neural network (CNN) architecture, dubbed HVPR, that integr ates both features into a single 3D representation effectively and efficiently. Specifically, we augment the point-based features with a memory module to reduce the computational cost. We then aggregate the features in the memory, semantica lly similar to each voxel-based one, to obtain a hybrid 3D representation in a f orm of a pseudo image, allowing to localize 3D objects in a single stage efficie

ntly. We also propose an Attentive Multi-scale Feature Module (AMFM) that extrac
ts scale-aware features considering the sparse and irregular patterns of point c
louds. Experimental results on the KITTI dataset demonstrate the effectiveness a
nd efficiency of our approach, achieving a better compromise in terms of speed a
nd accuracy.
****
SOE-Net: A Self-Attention and Orientation Encoding Network for Point Cloud Based
 Place Recognition
Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, Uwe Stilla; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2021, pp. 11348-11357
We tackle the problem of place recognition from point cloud data and introduce a
 self-attention and orientation encoding network (SOE-Net) that fully explores t
he relationship between points and incorporates long-range context into point-wi
se local descriptors. Local information of each point from eight orientations is
 captured in a PointOE module, whereas long-range feature dependencies among loc
al descriptors are captured with a self-attention unit. Moreover, we propose a n
ovel loss function called Hard Positive Hard Negative quadruplet loss (HPHN quad
ruplet), that achieves better performance than the commonly used metric learning
 loss. Experiments on various benchmark datasets demonstrate superior performanc
e of the proposed network over the current state-of-the-art approaches. Our code
 is released publicly at https://github.com/Yan-Xia/SOE-Net.
****
Controlling the Rain: From Removal to Rendering
Siqi Ni, Xueyun Cao, Tao Yue, Xuemei Hu; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6328-6337
Existing rain image editing methods focus on either removing rain from rain imag
es or rendering rain on rain-free images. This paper proposes to realize continu
ous control of rain intensity bidirectionally, from clear rain-free to downpour
image with a single rain image as input, without changing the scene-specific cha
racteristics, e.g. the direction, appearance and distribution of rain. Specifica
lly, we introduce a Rain Intensity Controlling Network (RICNet) that contains th
ree sub-networks of background extraction network, high-frequency rain-streak el
imination network and main controlling network, which allows to control rain ima
ge of different intensities continuously by interpolation in the deep feature sp
ace. The HOG loss and autocorrelation loss are proposed to enhance consistency i
n orientation and suppress repetitive rain streaks. Furthermore, a decremental l
earning strategy that trains the network from downpour to drizzle images sequent
ially is proposed to further improve the performance and speedup the convergence
. Extensive experiments on both rain dataset and real rain images demonstrate th
e effectiveness of the proposed method.
****
KeypointDeformer: Unsupervised 3D Keypoint Discovery for Shape Control
Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, Angjoo Kan
azawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2021, pp. 12783-12792
We introduce KeypointDeformer, a novel unsupervised method for shape control thr
ough automatically discovered 3D keypoints. We cast this as the problem of align
ing a source 3D object to a target 3D object from the same object category. Our
method analyzes the difference between the shapes of the two objects by comparin
g their latent representations. This latent representation is in the form of 3D
keypoints that are learned in an unsupervised way. The difference between the 3D
 keypoints of the source and the target objects then informs the shape deformati
on algorithm that deforms the source object into the target object. The whole mo
del is learned end-to-end and simultaneously discovers 3D keypoints while learni
ng to use them for deforming object shapes. Our approach produces intuitive and
semantically consistent control of shape deformations. Moreover, our discovered
3D keypoints are consistent across object category instances despite large shape
 variations. As our method is unsupervised, it can be readily deployed to new ob
ject categories without requiring annotations for 3D keypoints and deformations.

Project page: http://tomasjakab.github.io/KeypointDeformer
****
A2-FPN: Attention Aggregation Based Feature Pyramid Network for Instance Segmentation
Miao Hu, Yali Li, Lu Fang, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15343-15352
Learning pyramidal feature representations is crucial for recognizing object instances at different scales. Feature Pyramid Network (FPN) is the classic architecture to build a feature pyramid with high-level semantics throughout. However, intrinsic defects in feature extraction and fusion inhibit FPN from further aggregating more discriminative features. In this work, we propose Attention Aggregation based Feature Pyramid Network (A^2-FPN), to improve multi-scale feature learning through attention-guided feature aggregation. In feature extraction, it extracts discriminative features by collecting-distributing multi-level global context features, and mitigates the semantic information loss due to drastically reduced channels. In feature fusion, it aggregates complementary information from adjacent features to generate location-wise reassembly kernels for content-aware sampling, and employs channel-wise reweighting to enhance the semantic consistency before element-wise addition. A^2-FPN shows consistent gains on different instance segmentation frameworks. By replacing FPN with A^2-FPN in Mask R-CNN, our model boosts the performance by 2.1% and 1.6% mask AP when using ResNet-50 and ResNet-101 as backbone, respectively. Moreover, A^2-FPN achieves an improvement of 2.0% and 1.4% mask AP when integrated into the strong baselines such as Cascade Mask R-CNN and Hybrid Task Cascade.
****
Quasi-Dense Similarity Learning for Multiple Object Tracking
Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 164-173
Similarity learning has been recognized as a crucial step for object tracking. However, existing multiple object tracking methods only use sparse ground truth matching as the training objective, while ignoring the majority of the informative regions on the images. In this paper, we present Quasi-Dense Similarity Learning, which densely samples hundreds of region proposals on a pair of images for contrastive learning. We can directly combine this similarity learning with existing detection methods to build Quasi-Dense Tracking (QDTrack) without turning to displacement regression or motion priors. We also find that the resulting distinctive feature space admits a simple nearest neighbor search at the inference time. Despite its simplicity, QDTrack outperforms all existing methods on MOT, BDD100K, Waymo, and TAO tracking benchmarks. It achieves 68.7 MOTA at 20.3 FPS on MOT17 without using external training data. Compared to methods with similar detectors, it boosts almost 10 points of MOTA and significantly decreases the number of ID switches on BDD100K and Waymo datasets. Our code and trained models are available at https://github.com/SysCV/qdtrack.
****
Simultaneously Localize, Segment and Rank the Camouflaged Objects
Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, Deng-Ping Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11591-11601
Camouflage is a key defence mechanism across species that is critical to survival. Common camouflage include background matching, imitating the color and pattern of the environment, and disruptive coloration, disguising body outlines. Camouflaged object detection (COD) aims to segment camouflaged objects hiding in their surroundings. Existing COD models are built upon binary ground truth to segment the camouflaged objects without illustrating the level of camouflage. In this paper, we revisit this task and argue that explicitly modeling the conspicuousness of camouflaged objects against their particular backgrounds can not only lead to a better understanding about camouflage and evolution of animals, but also provide guidance to design more sophisticated camouflage techniques. Furthermore, we observe that it is some specific parts of the camouflaged objects that make

them detectable by predators. With the above understanding about camouflaged objects, we present the first ranking based COD network to simultaneously localize, segment and rank camouflaged objects. The localization model is proposed to find the discriminative regions that make the camouflaged object obvious. The segmentation model segments the full scope of the camouflaged objects. And, the ranking model infers the detectability of different camouflaged objects. Moreover, we contribute a large COD testing set to evaluate the generalization ability of COD models. Experimental results show that our model achieves new state-of-the-art, leading to a more interpretable COD network.
****
Hybrid Message Passing With Performance-Driven Structures for Facial Action Unit Detection
Tengfei Song, Zijun Cui, Wenming Zheng, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6267-6276
Message passing neural network has been an effective method to represent dependencies among nodes by propagating messages. However, most of message passing algorithms focus on one structure and the messages are estimated by one single approach. For the real-world data, like facial action units (AUs), the dependencies may vary in terms of different expressions and individuals. In this paper, we propose a novel hybrid message passing neural network with performance-driven structures (HMP-PS), which combines complementary message passing methods and captures more possible structures in a Bayesian manner. Particularly, a performance-driven Monte Carlo Markov Chain sampling method is proposed for generating high performance graph structures. Besides, the hybrid message passing is proposed to combine different types of messages, which provide the complementary information. The contribution of each type of message is adaptively adjusted along with different inputs. The experiments on two widely used benchmark datasets, i.e., BP4D and DISFA, validate that our proposed method can achieve the state-of-the-art performance.
****
Distilling Object Detectors via Decoupled Features
Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2154-2164
Knowledge distillation is a widely used paradigm for inheriting information from a complicated teacher network to a compact student network and maintaining the strong performance. Different from image classification, object detectors are much more sophisticated with multiple loss functions in which features that semantic information rely on are tangled. In this paper, we point out that the information of features derived from regions excluding objects are also essential for distilling the student detector, which is usually ignored in existing approaches. In addition, we elucidate that features from different regions should be assigned with different importance during distillation. To this end, we present a novel distillation algorithm via decoupled features (DeFeat) for learning a better student detector. Specifically, two levels of decoupled features will be processed for embedding useful information into the student, i.e., decoupled features from neck and decoupled proposals from classification head. Extensive experiments on various detectors with different backbones show that the proposed DeFeat is able to surpass the state-of-the-art distillation methods for object detection. For example, DeFeat improves ResNet50 based Faster R-CNN from 37.4% to 40.9% mAP, and improves ResNet50 based RetinaNet from 36.5% to 39.7% mAP on COCO benchmark. Code will be released.
****
Roof-GAN: Learning To Generate Roof Geometry and Relations for Residential Houses
Yiming Qian, Hao Zhang, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2796-2805
This paper presents Roof-GAN, a novel generative adversarial network that generates structured geometry of residential roof structures as a set of roof primitives and their relationships. Given the number of primitives, the generator produc

es a structured roof model as a graph, which consists of 1) primitive geometry as raster images at each node, encoding facet segmentation and angles; 2) inter-primitive colinear/coplanar relationships at each edge; and 3) primitive geometry in a vector format at each node, generated by a novel differentiable vectorizer while enforcing the relationships. The discriminator is trained to assess the primitive raster geometry, the primitive relationships, and the primitive vector geometry in a fully end-to-end architecture. Qualitative and quantitative evaluations demonstrate the effectiveness of our approach in generating diverse and realistic roof models over the competing methods with a novel metric proposed in this paper for the task of structured geometry generation. Code and data are available at https://github.com/yi-ming-qian/roofgan.
****
No Shadow Left Behind: Removing Objects and Their Shadows Using Approximate Lighting and Geometry
Edward Zhang, Ricardo Martin-Brualla, Janne Kontkanen, Brian L. Curless; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16397-16406
Removing objects from images is a challenging technical problem that is important for many applications, including mixed reality. For believable results, the shadows that the object casts should also be removed. Current inpainting-based methods only remove the object itself, leaving shadows behind, or at best require specifying shadow regions to inpaint. We introduce a deep learning pipeline for removing a shadow along with its caster. We leverage rough scene models in order to remove a wide variety of shadows (hard or soft, dark or subtle, large or thin) from surfaces with a wide variety of textures. We train our pipeline on synthetically rendered data, and show qualitative and quantitative results on both synthetic and real scenes.
****
NetAdaptV2: Efficient Neural Architecture Search With Fast Super-Network Training and Architecture Optimization
Tien-Ju Yang, Yi-Lun Liao, Vivienne Sze; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2402-2411
Neural architecture search (NAS) typically consists of three main steps: training a super-network, training and evaluating sampled deep neural networks (DNNs), and training the discovered DNN. Most of the existing efforts speed up some steps at the cost of a significant slowdown of other steps or sacrificing the support of non-differentiable search metrics. The unbalanced reduction in the time spent per step limits the total search time reduction, and the inability to support non-differentiable search metrics limits the performance of discovered DNNs. In this paper, we present NetAdaptV2 with three innovations to better balance the time spent for each step while supporting non-differentiable search metrics. First, we propose channel-level bypass connections that merge network depth and layer width into a single search dimension to reduce the time for training and evaluating sampled DNNs. Second, ordered dropout is proposed to train multiple DNNs in a single forward-backward pass to decrease the time for training a super-network. Third, we propose the multi-layer coordinate descent optimizer that considers the interplay of multiple layers in each iteration of optimization to improve the performance of discovered DNNs while supporting non-differentiable search metrics. With these innovations, NetAdaptV2 reduces the total search time by up to 5.8x on ImageNet and 2.4x on NYU Depth V2, respectively, and discovers DNNs with better accuracy-latency/accuracy-MAC trade-offs than state-of-the-art NAS works. Moreover, the discovered DNN outperforms NAS-discovered MobileNetV3 by 1.8% higher top-1 accuracy with the same latency.
****
PhD Learning: Learning With Pompeiu-Hausdorff Distances for Video-Based Vehicle Re-Identification
Jianan Zhao, Fengliang Qi, Guangyu Ren, Lin Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2225-2235
Vehicle re-identification (re-ID) is of great significance to urban operation, management, security and has gained more attention in recent years. However, two

critical challenges in vehicle re-ID have primarily been underestimated, i.e., 1
): how to make full use of raw data, and 2): how to learn a robust re-ID model w
ith noisy data. In this paper, we first create a video vehicle re-ID evaluation
benchmark called VVeRI-901 and verify the performance of video-based re-ID is fa
r better than static image-based one. Then we propose a new Pompeiu-hausdorff di
stance (PhD) learning method for video-to-video matching. It can alleviate the d
ata noise problem caused by the occlusion in videos and thus improve re-ID perfo
rmance significantly. Extensive empirical results on video-based vehicle and per
son re-ID datasets, i.e., VVeRI-901, MARS and PRID2011, demonstrate the superior
ity of the proposed method. The source code of our proposed method is available
at https://github.com/emdata-ailab/PhD-Learning.
****
DeepVideoMVS: Multi-View Stereo on Video With Recurrent Spatio-Temporal Fusion
Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu,
 Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2021, pp. 15324-15333
We propose an online multi-view depth prediction approach on posed video streams
, where the scene geometry information computed in the previous time steps is pr
opagated to the current time step in an efficient and geometrically plausible wa
y. The backbone of our approach is a real-time capable, lightweight encoder-deco
der that relies on cost volumes computed from pairs of images. We extend it by p
lacing a ConvLSTM cell at the bottleneck layer, which compresses an arbitrary am
ount of past information in its states. The novelty lies in propagating the hidd
en state of the cell by accounting for the viewpoint changes between time steps.
 At a given time step, we warp the previous hidden state into the current camera
 plane using the previous depth prediction. Our extension brings only a small ov
erhead of computation time and memory consumption, while improving the depth pre
dictions significantly. As a result, we outperform the existing state-of-the-art
 multi-view stereo methods on most of the evaluated metrics in hundreds of indoo
r scenes while maintaining a real-time performance. Code available: https://gith
ub.com/ardaduz/deep-video-mvs
****
Saliency-Guided Image Translation
Lai Jiang, Mai Xu, Xiaofei Wang, Leonid Sigal; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16509-16518
In this paper, we propose a novel task for saliency-guided image translation, wi
th the goal of image-to-image translation conditioned on the user specified sali
ency map. To address this problem, we develop a novel Generative Adversarial Net
work (GAN)-based model, called SalG-GAN. Given the original image and target sal
iency map, SalG-GAN can generate a translated image that satisfies the target sa
liency map. In SalG-GAN, a disentangled representation framework is proposed to
encourage the model to learn diverse translations for the same target saliency c
ondition. A saliency-based attention module is introduced as a special attention
 mechanism for facilitating the developed structures of saliency-guided generato
r, saliency cue encoder and saliency-guided global and local discriminators. Fur
thermore, we build a synthetic dataset and a real-world dataset with labeled vis
ual attention for training and evaluating our SalG-GAN. The experimental results
 over both datasets verify the effectiveness of our model for saliency-guided im
age translation.
****
Weakly Supervised Learning of Rigid 3D Scene Flow
Zan Gojcic, Or Litany, Andreas Wieser, Leonidas J. Guibas, Tolga Birdal; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2021, pp. 5692-5703
We propose a data-driven scene flow estimation algorithm exploiting the observat
ion that many 3D scenes can be explained by a collection of agents moving as rig
id bodies. At the core of our method lies a deep architecture able to reason at
the object-level by considering 3D scene flow in conjunction with other 3D tasks
. This object level abstraction enables us to relax the requirement for dense sc
ene flow supervision with simpler binary background segmentation mask and ego-mo

tion annotations. Our mild supervision requirements make our method well suited for recently released massive data collections for autonomous driving, which do not contain dense scene flow annotations. As output, our model provides low-level cues like pointwise flow and higher-level cues such as holistic scene understanding at the level of rigid objects. We further propose a test-time optimization refining the predicted rigid scene flow. We showcase the effectiveness and generalization capacity of our method on four different autonomous driving datasets. We release our source code and pre-trained models under github.com/zgojcic/Rigid3DSceneFlow.
****
InverseForm: A Loss Function for Structured Boundary-Aware Segmentation
Shubhankar Borse, Ying Wang, Yizhe Zhang, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5901-5911
We present a novel boundary-aware loss term for semantic segmentation using an inverse-transformation network, which efficiently learns the degree of parametric transformations between estimated and target boundaries. This plug-in loss term complements the cross-entropy loss in capturing boundary transformations and allows consistent and significant performance improvement on segmentation backbone models without increasing their size and computational complexity. We analyze the quantitative and qualitative effects of our loss function on three indoor and outdoor segmentation benchmarks, including Cityscapes, NYU-Depth-v2, and PASCAL, integrating it into the training phase of several backbone networks in both single-task and multi-task settings. Our extensive experiments show that the proposed method consistently outperforms baselines, and even sets the new state-of-the-art on two datasets.
****
Towards Accurate Text-Based Image Captioning With Content Diversity Exploration
Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12637-12646
Text-based image captioning (TextCap) which aims to read and reason images with texts is crucial for a machine to understand a detailed and complex scene environment, considering that texts are omnipresent in daily life. This task, however, is very challenging because an image often contains complex texts and visual information that is hard to be described comprehensively. Existing methods attempt to extend the traditional image captioning methods to solve this task, which focus on describing the overall scene of images by one global caption. This is infeasible because the complex text and visual information cannot be described well within one caption. To resolve this difficulty, we seek to generate multiple captions that accurately describe different parts of an image in detail. To achieve this purpose, there are three key challenges: 1) it is hard to decide which parts of the texts of images to copy or paraphrase; 2) it is non-trivial to capture the complex relationship between diverse texts in an image; 3) how to generate multiple captions with diverse content is still an open problem. To conquer these, we propose a novel Anchor-Captioner method. Specifically, we first find the important tokens which are supposed to be paid more attention to and consider them as anchors. Then, for each chosen anchor, we group its relevant texts to construct the corresponding anchor-centred graph (ACG). Last, based on different ACGs, we conduct the multi-view caption generation to improve the content diversity of generated captions. Experimental results show that our method not only achieves SOTA performance but also generates diverse captions to describe images.
****
Learning Placeholders for Open-Set Recognition
Da-Wei Zhou, Han-Jia Ye, De-Chuan Zhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4401-4410
Traditional classifiers are deployed under closed-set setting, with both training and test classes belong to the same set. However, real-world applications probably face the input of unknown categories, and the model will recognize them as known ones. Under such circumstances, open-set recognition is proposed to mainta

in classification performance on known classes and reject unknowns. The closed-set models make overconfident predictions over familiar known class instances, so that calibration and thresholding across categories become essential issues when extending to an open-set environment. To this end, we proposed to learn PlaceholdeRs for Open-SEt Recognition (Proser), which prepares for the unknown classes by allocating placeholders for both data and classifier. In detail, learning data placeholders tries to anticipate open-set class data, thus transforms closed-set training into open-set training. Besides, to learn the invariant information between target and non-target classes, we reserve classifier placeholders as the class-specific boundary between known and unknown. The proposed Proser efficiently generates novel class by manifold mixup, and adaptively sets the value of reserved open-set classifier during training. Experiments on various datasets validate the effectiveness of our proposed method.
****
CodedStereo: Learned Phase Masks for Large Depth-of-Field Stereo
Shiyu Tan, Yicheng Wu, Shoou-I Yu, Ashok Veeraraghavan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7170-7179
Conventional stereo suffers from a fundamental trade-off between imaging volume and signal-to-noise ratio (SNR) -- due to the conflicting impact of aperture size on both these variables. Inspired by the extended depth of field cameras, we propose a novel end-to-end learning-based technique to overcome this limitation, by introducing a phase mask at the aperture plane of the cameras in a stereo imaging system. The phase mask creates a depth-dependent point spread function, allowing us to recover sharp image texture and stereo correspondence over a significantly extended depth of field (EDOF) than conventional stereo. The phase mask pattern, the EDOF image reconstruction, and the stereo disparity estimation are all trained together using an end-to-end learned deep neural network. We perform theoretical analysis and characterization of the proposed approach and show a 6x increase in volume that can be imaged in simulation. We also build an experimental prototype and validate the approach using real-world results acquired using this prototype system.
****
More Photos Are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval
Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4247-4256
A fundamental challenge faced by existing Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) models is the data scarcity -- model performances are largely bottlenecked by the lack of sketch-photo pairs. Whilst the number of photos can be easily scaled, each corresponding sketch still needs to be individually produced. In this paper, we aim to mitigate such an upper-bound on sketch data, and study whether unlabelled photos alone (of which they are many) can be cultivated for performance gain. In particular, we introduce a novel semi-supervised framework for cross-modal retrieval that can additionally leverage large-scale unlabelled photos to account for data scarcity. At the center of our semi-supervision design is a sequential photo-to-sketch generation model that aims to generate paired sketches for unlabelled photos. Importantly, we further introduce a discriminator-guided mechanism to guide against unfaithful generation, together with a distillation loss-based regularizer to provide tolerance against noisy training samples. Last but not least, we treat generation and retrieval as two conjugate problems, where a joint learning procedure is devised for each module to mutually benefit from each other. Extensive experiments show that our semi-supervised model yields a significant performance boost over the state-of-the-art supervised alternatives, as well as existing methods that can exploit unlabelled photos for FG-SBIR.
****
Unsupervised Hyperbolic Representation Learning via Message Passing Auto-Encoders

Jiwoong Park, Junho Cho, Hyung Jin Chang, Jin Young Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5516-5526
Most of the existing literature regarding hyperbolic embedding concentrate upon supervised learning, whereas the use of unsupervised hyperbolic embedding is less well explored. In this paper, we analyze how unsupervised tasks can benefit from learned representations in hyperbolic space. To explore how well the hierarchical structure of unlabeled data can be represented in hyperbolic spaces, we design a novel hyperbolic message passing auto-encoder whose overall auto-encoding is performed in hyperbolic space. The proposed model conducts auto-encoding the networks via fully utilizing hyperbolic geometry in message passing. Through extensive quantitative and qualitative analyses, we validate the properties and benefits of the unsupervised hyperbolic representations. Codes are available at https://github.com/junhocho/HGCAE.
****
Retinex-Inspired Unrolling With Cooperative Prior Architecture Search for Low-Light Image Enhancement
Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, Zhongxuan Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10561-10570
Low-light image enhancement plays very important roles in low-level vision areas. Recent works have built a great deal of deep learning models to address this task. However, these approaches mostly rely on significant architecture engineering and suffer from high computational burden. In this paper, we propose a new method, named Retinex-inspired Unrolling with Architecture Search (RUAS), to construct lightweight yet effective enhancement network for low-light images in real-world scenario. Specifically, building upon Retinex rule, RUAS first establishes models to characterize the intrinsic underexposed structure of low-light images and unroll their optimization processes to construct our holistic propagation structure. Then by designing a cooperative reference-free learning strategy to discover low-light prior architectures from a compact search space, RUAS is able to obtain a top-performing image enhancement network, which is with fast speed and requires few computational resources. Extensive experiments verify the superiority of our RUAS framework against recently proposed state-of-the-art methods. The project page is available at http://dutmedia.org/RUAS/.
****
Relevance-CAM: Your Model Already Knows Where To Look
Jeong Ryong Lee, Sewon Kim, Inyong Park, Taejoon Eo, Dosik Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14944-14953
With increasing fields of application for neural networks and the development of neural networks, the ability to explain deep learning models is also becoming increasingly important. Especially, prior to practical applications, it is crucial to analyze a model's inference and the process of generating the results. A common explanation method is Class Activation Mapping(CAM) based method where it is often used to understand the last layer of the convolutional neural networks popular in the field of Computer Vision. In this paper, we propose a novel CAM method named Relevance-weighted Class Activation Mapping(Relevance-CAM) that utilizes Layer-wise Relevance Propagation to obtain the weighting components. This allows the explanation map to be faithful and robust to the shattered gradient problem, a shared problem of the gradient based CAM methods that causes noisy saliency maps for intermediate layers. Therefore, our proposed method can better explain a model by correctly analyzing the intermediate layers as well as the last convolutional layer. In this paper, we visualize how each layer of the popular image processing models extracts class specific features using Relevance-CAM, evaluate the localization ability, and show why the gradient based CAM cannot be used to explain the intermediate layers, proven by experimenting the weighting component. Relevance-CAM outperforms other CAM-based methods in recognition and localization evaluation in layers of any depth. The source code is available at: https://github.com/mongeoroo/Relevance-CAM

****
Boundary IoU: Improving Object-Centric Image Segmentation Evaluation
Bowen Cheng, Ross Girshick, Piotr Dollar, Alexander C. Berg, Alexander Kirillov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15334-15342
We present Boundary IoU (Intersection-over-Union), a new segmentation evaluation measure focused on boundary quality. We perform an extensive analysis across different error types and object sizes and show that Boundary IoU is significantly more sensitive than the standard Mask IoU measure to boundary errors for large objects and does not over-penalize errors on smaller objects. The new quality measure displays several desirable characteristics like symmetry w.r.t. prediction/ground truth pairs and balanced responsiveness across scales, which makes it more suitable for segmentation evaluation than other boundary-focused measures like Trimap IoU and F-measure. Based on Boundary IoU, we update the standard evaluation protocols for instance and panoptic segmentation tasks by proposing the Boundary AP (Average Precision) and Boundary PQ (Panoptic Quality) metrics, respectively. Our experiments show that the new evaluation metrics track boundary quality improvements that are generally overlooked by current Mask IoU-based evaluation metrics. We hope that the adoption of the new boundary-sensitive evaluation metrics will lead to rapid progress in segmentation methods that improve boundary quality.
****
KeepAugment: A Simple Information-Preserving Data Augmentation Approach
Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, Qiang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1055-1064
Data augmentation (DA) is an essential technique for training state-of-the-art deep learning systems. In this paper, we empirically show data augmentation might introduce noisy augmented examples and consequently hurt the performance on unaugmented data during inference. To alleviate this issue, we propose a simple yet highly effective approach, dubbed KeepAugment, to increase augmented images fidelity. The idea is first to use the saliency map to detect important regions on the original images and then preserve these informative regions during augmentation. This information-preserving strategy allows us to generate more faithful training examples. Empirically, we demonstrate our method significantly improves on a number of prior art data augmentation schemes, e.g. AutoAugment, Cutout, random erasing, achieving promising results on image classification, semi-supervised image classification, multi-view multi-camera tracking and object detection.
****
On Robustness and Transferability of Convolutional Neural Networks
Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, Mario Lucic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16458-16468
Modern deep convolutional networks (CNNs) are often criticized for not generalizing under distributional shifts. However, several recent breakthroughs in transfer learning suggest that these networks can cope with severe distribution shifts and successfully adapt to new tasks from a few training examples. In this work we study the interplay between out-of-distribution and transfer performance of modern image classification CNNs for the first time and investigate the impact of the pre-training data size, the model scale, and the data preprocessing pipeline. We find that increasing both the training set and model sizes significantly improve the distributional shift robustness. Furthermore, we show that, perhaps surprisingly, simple changes in the preprocessing such as modifying the image resolution can significantly mitigate robustness issues in some cases. Finally, we outline the shortcomings of existing robustness evaluation datasets and introduce a synthetic dataset SI-Score we use for a systematic analysis across factors of variation common in visual data such as object scale and position.
****

POSEFusion: Pose-Guided Selective Fusion for Single-View Human Volumetric Capture
Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14162-14172
We propose POse-guided SElective Fusion (POSEFusion), a single-view human volumetric capture method that leverages tracking-based methods and tracking-free inference to achieve high-fidelity and dynamic 3D reconstruction. By contributing a novel reconstruction framework which contains pose-guided keyframe selection and robust implicit surface fusion, our method fully utilizes the advantages of both tracking-based methods and tracking-free inference methods, and finally enables the high-fidelity reconstruction of dynamic surface details even in the invisible regions. We formulate the keyframe selection as a dynamic programming problem to guarantee the temporal continuity of the reconstructed sequence. Moreover, the novel robust implicit surface fusion involves an adaptive blending weight to preserve high-fidelity surface details and an automatic collision handling method to deal with the potential self-collisions. Overall, our method enables high-fidelity and dynamic capture in both visible and invisible regions from a single RGBD camera, and the results and experiments show that our method outperforms state-of-the-art methods.
****
Exploring Adversarial Fake Images on Face Manifold
Dongze Li, Wei Wang, Hongxing Fan, Jing Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5789-5798
Images synthesized by powerful generative adversarial network (GAN) based methods have drawn moral and privacy concerns. Although image forensic models have reached great performance in detecting fake images from real ones, these models can be easily fooled with a simple adversarial attack. But, the noise adding adversarial samples are also arousing suspicion. In this paper, instead of adding adversarial noise, we optimally search adversarial points on face manifold to generate anti-forensic fake face images. We iteratively do a gradient-descent with each small step in the latent space of a generative model, e.g. Style-GAN, to find an adversarial latent vector, which is similar to norm-based adversarial attack but in latent space. Then, the generated fake images driven by the adversarial latent vectors with the help of GANs can defeat main-stream forensic models. For examples, they make the accuracy of deepfake detection models based on Xception or EfficientNet drop from over 90% to nearly 0%, meanwhile maintaining high visual quality. In addition, we find manipulating noise vectors n at different levels have different impacts on attack success rate, and the generated adversarial images mainly have changes on facial texture or face attributes.
****
Reinforced Attention for Few-Shot Learning and Beyond
Jie Hong, Pengfei Fang, Weihao Li, Tong Zhang, Christian Simon, Mehrtash Harandi, Lars Petersson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 913-923
Few-shot learning aims to correctly recognize query samples from unseen classes given a limited number of support samples, often by relying on global embeddings of images. In this paper, we propose to equip the backbone network with an attention agent, which is trained by reinforcement learning. The policy gradient algorithm is employed to train the agent towards adaptively localizing the representative regions on feature maps over time. We further design a reward function based on the prediction of the held-out data, thus helping the attention mechanism to generalize better across the unseen classes. The extensive experiments show, with the help of the reinforced attention, that our embedding network has the capability to progressively generate a more discriminative representation in few-shot learning. Moreover, experiments on the task of image classification also show the effectiveness of the proposed design.
****
HOTR: End-to-End Human-Object Interaction Detection With Transformers
Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, Hyunwoo J. Kim; Proceedings o

f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 74-83
Human-Object Interaction (HOI) detection is a task of identifying "a set of inte
ractions" in an image, which involves the i) localization of the subject (i.e.,
humans) and target (i.e., objects) of interaction, and ii) the classification of
 the interaction labels. Most existing methods have addressed this task in an in
direct way by detecting human and object instances and individually inferring ev
ery pair of the detected instances. In this paper, we present a novel framework,
 referred by HOTR, which directly predicts a set of <human, object, interaction>
 triplets from an image based on a transformer encoder-decoder architecture. Thr
ough the set prediction, our method effectively exploits the inherent semantic r
elationships in an image and does not require time-consuming post-processing whi
ch is the main bottleneck of existing methods. Our proposed algorithm achieves t
he state-of-the-art performance in two HOI detection benchmarks with an inferenc
e time under 1 ms after object detection.
****
Deep Video Matting via Spatio-Temporal Alignment and Aggregation
Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, Yu-Wing Tai; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p
p. 6975-6984
Despite the significant progress made by deep learning in natural image matting,
 there has been so far no representative work on deep learning for video matting
 due to the inherent technical challenges in reasoning temporal domain and lack
of large-scale video matting datasets. In this paper, we propose a deep learning
-based video matting framework which employs a novel and effective spatio-tempor
al feature aggregation module (ST-FAM). As optical flow estimation can be very u
nreliable within matting regions, ST-FAM is designed to effectively align and ag
gregate information across different spatial scales and temporal frames within t
he network decoder. To eliminate frame-by-frame trimap annotations, a lightweigh
t interactive trimap propagation network is also introduced. The other contribut
ion consists of a large-scale video matting dataset with groundtruth alpha matte
s for quantitative evaluation and real-world high-resolution videos with trimaps
 for qualitative evaluation. Quantitative and qualitative experimental results s
how that our framework significantly outperforms conventional video matting and
deep image matting methods applied to video in presence of multi-frame temporal
information.
****
Triple-Cooperative Video Shadow Detection
Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, Jing Qin; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2021, pp. 2715-2724
Shadow detection in single image has received signifi-cant research interests in
 recent years. However, much lessworks has been explored in shadow detection ove
r dynamicscenes. The bottleneck is the lack of a well-establisheddataset with hi
gh-quality annotations for video shadow de-tection. In this work, we collect a n
ew video shadow detec-tion dataset (ViSha), which contains120videos with11,685fr
ames, covering 60 object categories, varying lengths, anddifferent motion/lighti
ng conditions. All the frames are an-notated with a high-quality pixel-level sha
dow mask. Tothe best of our knowledge, this is the first learning-orienteddatase
t for video shadow detection. Furthermore, we de-velop a new baseline model, nam
ed triple-cooperative videoshadow detection network (TVSD-Net). It utilizes trip
leparallel networks in a cooperative manner to learn discrim-inative representat
ions at intra-video and inter-video lev-els. Within the network, a dual gated co
-attention moduleis proposed to constrain features from neighboring framesin the
 same video, while an auxiliary similarity loss is in-troduced to mine semantic
information between differentvideos. Finally, we conduct a comprehensive study o
n ViShadataset, systematically evaluating 10 state-of-the-art mod-els (including
 single image shadow detectors, video ob-ject and saliency detection methods). E
xperimental resultsdemonstrate that our model outperforms SOTA competitors.
****

Scale-Aware Graph Neural Network for Few-Shot Semantic Segmentation
Guo-Sen Xie, Jie Liu, Huan Xiong, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5475-5484
Few-shot semantic segmentation (FSS) aims to segment unseen class objects given very few densely-annotated support images from the same class. Existing FSS methods find the query object by using support prototypes or by directly relying on heuristic multi-scale feature fusion. However, they fail to fully leverage the high-order appearance relationships between multi-scale features among the support-query image pairs, thus leading to an inaccurate localization of the query objects. To tackle the above challenge, we propose an end-to-end scale-aware graph neural network (SAGNN) by reasoning the cross-scale relations among the support-query images for FSS. Specifically, a scale-aware graph is first built by taking support-induced multi-scale query features as nodes and, meanwhile, each edge is modeled as the pairwise interaction of its connected nodes. By progressive message passing over this graph, SAGNN is capable of capturing cross-scale relations and overcoming object variations (e.g., appearance, scale and location), and can thus learn more precise node embeddings. This in turn enables it to predict more accurate foreground objects. Moreover, to make full use of the location relations across scales for the query image, a novel self-node collaboration mechanism is proposed to enrich the current node, which endows SAGNN the ability of perceiving different resolutions of the same objects. Extensive experiments on PASCAL-5i and COCO-20i show that SAGNN achieves state-of-the-art results.
****
Continuous Face Aging via Self-Estimated Residual Age Embedding
Zeqi Li, Ruowei Jiang, Parham Aarabi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15008-15017
Face synthesis, including face aging, in particular, has been one of the major topics that witnessed a substantial improvement in image fidelity by using generative adversarial networks (GANs). Most existing face aging approaches divide the dataset into several age groups and leverage group-based training strategies, which lacks the ability to provide fine-controlled continuous aging synthesis in nature. In this work, we propose a unified network structure that embeds a linear age estimator into a GAN-based model, where the embedded age estimator is trained jointly with the encoder and decoder to estimate the age of a face image and provide a personalized target age embedding for age progression/regression. The personalized target age embedding is synthesized by incorporating both personalized residual age embedding of the current age and exemplar-face aging basis of the target age, where all preceding aging bases are derived from the learned weights of the linear age estimator. This formulation brings the unified perspective of estimating the age and generating personalized aged face, where self-estimated age embeddings can be learned for every single age. The qualitative and quantitative evaluations on different datasets further demonstrate the significant improvement in the continuous face aging aspect over the state-of-the-art.
****
Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline
Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9229-9238
Depth maps obtained by commercial depth sensors are always in low-resolution, making it difficult to be used in various computer vision tasks. Thus, depth map super-resolution (SR) is a practical and valuable task, which upscales the depth map into high-resolution (HR) space. However, limited by the lack of real-world paired low-resolution (LR) and HR depth maps, most existing methods use downsampling to obtain paired training samples. To this end, we first construct a large-scale dataset named "RGB-D-D", which can greatly promote the study of depth map SR and even more depth-related real-world tasks. The "D-D" in our dataset represents the paired LR and HR depth maps captured from mobile phone and Lucid Helios respectively ranging from indoor scenes to challenging outdoor scenes. Besides, we provide a fast depth map super-resolution (FDSR) baseline, in which the high

-frequency component adaptively decomposed from RGB image to guide the depth map SR. Extensive experiments on existing public datasets demonstrate the effectiveness and efficiency of our network compared with the state-of-the-art methods. Moreover, for the real-world LR depth maps, our algorithm can produce more accurate HR depth maps with clearer boundaries and to some extent correct the depth value errors.
****
Jigsaw Clustering for Unsupervised Visual Representation Learning
Pengguang Chen, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11526-11535
Unsupervised representation learning with contrastive learning achieves great success recently. However, these methods have to duplicate each training batch to construct contrastive pairs, ie, each training batch and its augmented version should be forwarded simultaneously, leading to nearly double computation resource demand. We propose a novel Jigsaw Clustering pretext task in this paper, which only needs to forward each training batch itself, nearly reducing the training cost by a half. Our method makes use of information from both intra-image and inter-images, and outperforms previous single-batch based methods by a large margin, even comparable to the costly contrastive learning methods with only half the number of training batches. Our method shows that multiple batches during training are not necessary, and opens a new door for future research of single-batch based unsupervised methods. Our models trained on ImageNet datasets achieve state-of-the-art results with linear classification, outperform previous single-batch methods by 2.6%. Models transfer to COCO datasets outperforms MoCo v2 by 0.4% with only half the number of training samples. Our pretrained models outperform supervised ImageNet pretrained models on CIFAR-10 and CIFAR-100 datasets by 0.9% and 4.1% respectively.
****
DI-Fusion: Online Implicit 3D Reconstruction With Deep Priors
Jiahui Huang, Shi-Sheng Huang, Haoxuan Song, Shi-Min Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8932-8941
Previous online 3D dense reconstruction methods struggle to achieve the balance between memory storage and surface quality, largely due to the usage of stagnant underlying geometry representation, such as TSDF (truncated signed distance functions) or surfels, without any knowledge of the scene priors. In this paper, we present DI-Fusion (Deep Implicit Fusion), based on a novel 3D representation, i.e. Probabilistic Local Implicit Voxels (PLIVoxs), for online 3D reconstruction with a commodity RGB-D camera. Our PLIVox encodes scene priors considering both the local geometry and uncertainty parameterized by a deep neural network. With such deep priors, we are able to perform online implicit 3D reconstruction achieving state-of-the-art camera trajectory estimation accuracy and mapping quality, while achieving better storage efficiency compared with previous online 3D reconstruction approaches.
****
Square Root Bundle Adjustment for Large-Scale Reconstruction
Nikolaus Demmel, Christiane Sommer, Daniel Cremers, Vladyslav Usenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11723-11732
We propose a new formulation for the bundle adjustment problem which relies on nullspace marginalization of landmark variables by QR decomposition. Our approach, which we call square root bundle adjustment, is algebraically equivalent to the commonly used Schur complement trick, improves the numeric stability of computations, and allows for solving large-scale bundle adjustment problems with single-precision floating-point numbers. We show in real-world experiments with the BAL datasets that even in single precision the proposed solver achieves on average equally accurate solutions compared to Schur complement solvers using double precision. It runs significantly faster, but can require larger amounts of memory on dense problems. The proposed formulation relies on simple linear algebra operations and opens the way for efficient implementations of bundle adjustment on

hardware platforms optimized for single-precision linear algebra processing.
****
PatchMatch-Based Neighborhood Consensus for Semantic Correspondence
Jae Yong Lee, Joseph DeGol, Victor Fragoso, Sudipta N. Sinha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13153-13163
We address estimating dense correspondences between two images depicting different but semantically related scenes. End-to-end trainable deep neural networks incorporating neighborhood consensus cues are currently the best methods for this task. However, these architectures require exhaustive matching and 4D convolutions over matching costs for all pairs of feature map pixels. This makes them computationally expensive. We present a more efficient neighborhood consensus approach based on PatchMatch. For higher accuracy, we propose to use a learned local 4D scoring function for evaluating candidates during the PatchMatch iterations. We have devised an approach to jointly train the scoring function and the feature extraction modules by embedding them into a proxy model which is end-to-end differentiable. The modules are trained in a supervised setting using a cross-entropy loss to directly incorporate sparse keypoint supervision. Our evaluation on PF-Pascal and SPair-71K shows that our method significantly outperforms the state-of-the-art on both datasets while also being faster and using less memory.
****
Representative Forgery Mining for Fake Face Detection
Chengrui Wang, Weihong Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14923-14932
Although vanilla Convolutional Neural Network (CNN) based detectors can achieve satisfactory performance on fake face detection, we observe that the detectors tend to seek forgeries on a limited region of face, which reveals that the detectors is short of understanding of forgery. Therefore, we propose an attention-based data augmentation framework to guide detector refine and enlarge its attention. Specifically, our method tracks and occludes the Top-N sensitive facial regions, encouraging the detector to mine deeper into the regions ignored before for more representative forgery. Especially, our method is simple-to-use and can be easily integrated with various CNN models. Extensive experiments show that the detector trained with our method is capable to separately point out the representative forgery of fake faces generated by different manipulation techniques, and our method enables a vanilla CNN-based detector to achieve state-of-the-art performance without structure modification.
****
Look Closer To Segment Better: Boundary Patch Refinement for Instance Segmentation
Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, Xiaolin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13926-13935
Tremendous efforts have been made on instance segmentation but the mask quality is still not satisfactory. The boundaries of predicted instance masks are usually imprecise due to the low spatial resolution of feature maps and the imbalance problem caused by the extremely low proportion of boundary pixels. To address these issues, we propose a conceptually simple yet effective post-processing refinement framework to improve the boundary quality based on the results of any instance segmentation model, termed BPR. Following the idea of looking closer to segment boundaries better, we extract and refine a series of small boundary patches along the predicted instance boundaries. The refinement is accomplished by a boundary patch refinement network at higher resolution. The proposed BPR framework yields significant improvements over the Mask R-CNN baseline on Cityscapes benchmark, especially on the boundary-aware metrics. Moreover, by applying the BPR framework to the PolyTransform + SegFix baseline, we reached 1st place on the Cityscapes leaderboard.
****
Adaptive Class Suppression Loss for Long-Tail Object Detection
Tong Wang, Yousong Zhu, Chaoyang Zhao, Wei Zeng, Jinqiao Wang, Ming Tang; Procee

dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3103-3112
To address the problem of long-tail distribution for the large vocabulary object detection task, existing methods usually divide the whole categories into several groups and treat each group with different strategies. These methods bring the following two problems. One is the training inconsistency between adjacent categories of similar sizes, and the other is that the learned model is lack of discrimination for tail categories which are semantically similar to some of the head categories. In this paper, we devise a novel Adaptive Class Suppression Loss (ACSL) to effectively tackle the above problems and improve the detection performance of tail categories. Specifically, we introduce a statistic-free perspective to analyze the long-tail distribution, breaking the limitation of manual grouping. According to this perspective, our ACSL adjusts the suppression gradients for each sample of each class adaptively, ensuring the training consistency and boosting the discrimination for rare categories. Extensive experiments on long-tail datasets LVIS and Open Images show that the our ACSL achieves 5.18% and 5.2% improvements with ResNet50-FPN, and sets a new state of the art. Code and models are available at https://github.com/CASIA-IVA-Lab/ACSL.
****
ChallenCap: Monocular 3D Capture of Challenging Human Performances Using Multi-Modal References
Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11400-11411
Capturing challenging human motions is critical for numerous applications, but it suffers from complex motion patterns and severe self-occlusion under the monocular setting. In this paper, we propose ChallenCap --- a template-based approach to capture challenging 3D human motions using a single RGB camera in a novel learning-and-optimization framework, with the aid of multi-modal references. We propose a hybrid motion inference stage with a generation network, which utilizes a temporal encoder-decoder to extract the motion details from the pair-wise sparse-view reference, as well as a motion discriminator to utilize the unpaired marker-based references to extract specific challenging motion characteristics in a data-driven manner. We further adopt a robust motion optimization stage to increase the tracking accuracy, by jointly utilizing the learned motion details from the supervised multi-modal references as well as the reliable motion hints from the input image reference. Extensive experiments on our new challenging motion dataset demonstrate the effectiveness and robustness of our approach to capture challenging human motions.
****
Automated Log-Scale Quantization for Low-Cost Deep Neural Networks
Sangyun Oh, Hyeonuk Sim, Sugil Lee, Jongeun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 742-751
Quantization plays an important role in deep neural network (DNN) hardware. In particular, logarithmic quantization has multiple advantages for DNN hardware implementations, and its weakness in terms of lower performance at high precision compared with linear quantization has been recently remedied by what we call selective two-word logarithmic quantization (STLQ). However, there is a lack of training methods designed for STLQ or even logarithmic quantization in general. In this paper we propose a novel STLQ-aware training method, which significantly outperforms the previous state-of-the-art training method for STLQ. Moreover, our training results demonstrate that with our new training method, STLQ applied to weight parameters of ResNet-18 can achieve the same level of performance as state-of-the-art quantization method, APoT, at 3-bit precision. We also apply our method to various DNNs in image enhancement and semantic segmentation, showing competitive results.
****
Hallucination Improves Few-Shot Object Detection
Weilin Zhang, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13008-13017

Learning to detect novel objects with a few instances is challenging. A particul
arly challenging but practical regime is the extremely-low-shot regime (less tha
n three training examples). One critical factor in improving few-shot detection
is to handle the lack of variation in training data. The classifier relies on hi
gh intersection-over-union (IOU) boxes reported by the RPN to build a model of t
he category's variation in appearance. With only a few training examples, the va
riations are insufficient to train the classifier in novel classes. We propose t
o build a better model of variation in novel classes by transferring the shared
within-class variation from base classes. We introduce a hallucinator network an
d insert it into a modern object detector model, which learns to generate additi
onal training examples in the Region of Interest (ROI's) feature space. Our appr
oach yields significant performance improvements on two state-of-the-art few-sho
t detectors with different proposal generation processes. We achieve new state-o
f-the-art in very low-shot regimes on widely used benchmarks PASCAL VOC and COCO
.
****
Efficient Conditional GAN Transfer With Knowledge Propagation Across Classes
Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2021, pp. 12167-12176
Generative adversarial networks (GANs) have shown impressive results in both unc
onditional and conditional image generation. In recent literature, it is shown t
hat pre-trained GANs, on a different dataset, can be transferred to improve the
image generation from a small target data. The same, however, has not been well-
studied in the case of conditional GANs (cGANs), which provides new opportunitie
s for knowledge transfer compared to unconditional setup. In particular, the new
 classes may borrow knowledge from the related old classes, or share knowledge a
mong themselves to improve the training. This motivates us to study the problem
of efficient conditional GAN transfer with knowledge propagation across classes.
 To address this problem, we introduce a new GAN transfer method to explicitly p
ropagate the knowledge from the old classes to the new classes. The key idea is
to enforce the popularly used conditional batch normalization (BN) to learn the
class-specific information of the new classes from that of the old classes, with
 implicit knowledge sharing among the new ones. This allows for an efficient kno
wledge propagation from the old classes to the new ones, with the BN parameters
increasing linearly with the number of new classes. The extensive evaluation dem
onstrates the clear superiority of the proposed method over state-of-the-art com
petitors for efficient conditional GAN transfer tasks. The code is available at:
 https://github.com/mshahbazi72/cGANTransfer
****
Fully Convolutional Scene Graph Generation
Hengyue Liu, Ning Yan, Masood Mortazavi, Bir Bhanu; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11546-11
556
This paper presents a fully convolutional scene graph generation (FCSGG) model t
hat detects objects and relations simultaneously. Most of the scene graph genera
tion frameworks use a pre-trained two-stage object detector, like Faster R-CNN,
and build scene graphs using bounding box features. Such pipeline usually has a
large number of parameters and low inference speed. Unlike these approaches, FCS
GG is a conceptually elegant and efficient bottom-up approach that encodes objec
ts as bounding box center points, and relationships as 2D vector fields which ar
e named as Relation Affinity Fields (RAFs). RAFs encode both semantic and spatia
l features, and explicitly represent the relationship between a pair of objects
by the integral on a sub-region that points from subject to object. FCSGG only u
tilizes visual features and still generates strong results for scene graph gener
ation. Comprehensive experiments on the Visual Genome dataset demonstrate the ef
ficacy, efficiency, and generalizability of the proposed method. FCSGG achieves
highly competitive results on recall and zero-shot recall with significantly red
uced inference time.
****

Crossing Cuts Polygonal Puzzles: Models and Solvers
Peleg Harel, Ohad Ben-Shahar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3084-3093
Jigsaw puzzle solving, the problem of constructing a coherent whole from a set of non-overlapping unordered fragments, is fundamental to numerous applications, and yet most of the literature has focused thus far on less realistic puzzles whose pieces are identical squares. Here we formalize a new type of jigsaw puzzle where the pieces are general convex polygons generated by cutting through a global polygonal shape with an arbitrary number of straight cuts. We analyze the theoretical properties of such puzzles, including the inherent challenges in solving them once pieces are contaminated with geometrical noise. To cope with such difficulties and obtain tractable solutions, we abstract the problem as a multi-body spring-mass dynamical system endowed with hierarchical loop constraints and a layered reconstruction process that is guided by the pictorial content of the pieces. We define evaluation metrics and present experimental results on both apictorial and pictorial puzzles to indicate that they are solvable completely automatically.
****
Graph-Based High-Order Relation Modeling for Long-Term Action Recognition
Jiaming Zhou, Kun-Yu Lin, Haoxin Li, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8984-8993
Long-term actions involve many important visual concepts, e.g., objects, motions, and sub-actions, and there are various relations among these concepts, which we call basic relations. These basic relations will jointly affect each other during the temporal evolution of long-term actions, which forms the high-order relations that are essential for long-term action recognition. In this paper, we propose a Graph-based High-order Relation Modeling (GHRM) module to exploit the high-order relations in the long-term actions for long-term action recognition. In GHRM, each basic relation in the long-term actions will be modeled by a graph, where each node represents a segment in a long video. Moreover, when modeling each basic relation, the information from all the other basic relations will be incorporated by GHRM, and thus the high-order relations in the long-term actions can be well exploited. To better exploit the high-order relations along the time dimension, we design a GHRM-layer consisting of a Temporal-GHRM branch and a Semantic-GHRM branch, which aims to model the local temporal high-order relations and global semantic high-order relations. The experimental results on three long-term action recognition datasets, namely, Breakfast, Charades, and MultiThumos, demonstrate the effectiveness of our model.
****
Positive-Unlabeled Data Purification in the Wild for Object Detection
Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Xinghao Chen, Chunjing Xu, Chang Xu, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2653-2662
Deep learning based object detection approaches have achieved great progress with the benefit from large amount of labeled images. However, image annotation remains a laborious, time-consuming and error-prone process. To further improve the performance of detectors, we seek to exploit all available labeled data and excavate useful samples from massive unlabeled images in the wild, which is rarely discussed before. In this paper, we present a positive-unlabeled learning based scheme to expand training data by purifying valuable images from massive unlabeled ones, where the original training data are viewed as positive data and the unlabeled images in the wild are unlabeled data. To effectively utilized these purified data, we propose a self-distillation algorithm based on hint learning and ground truth bounded knowledge distillation. Experimental results verify that the proposed positive-unlabeled data purification can strengthen the original detector by mining the massive unlabeled data. In particular, our method boosts the mAP of FPN by +2.0% on COCO benchmark.
****
ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows

Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 862-871
Universal style transfer retains styles from reference images in content images. While existing methods have achieved state-of-the-art style transfer performance, they are not aware of the content leak phenomenon that the image content may corrupt after several rounds of stylization process. In this paper, we propose ArtFlow to prevent content leak during universal style transfer. ArtFlow consists of reversible neural flows and an unbiased feature transfer module. It supports both forward and backward inferences and operates in a projection-transfer-reversion scheme. The forward inference projects input images into deep features, while the backward inference remaps deep features back to input images in a lossless and unbiased way. Extensive experiments demonstrate that ArtFlow achieves comparable performance to state-of-the-art style transfer methods while avoiding content leak.
****
Network Quantization With Element-Wise Gradient Scaling
Junghyup Lee, Dohyung Kim, Bumsub Ham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6448-6457
Network quantization aims at reducing bit-widths of weights and/or activations, particularly important for implementing deep neural networks with limited hardware resources. Most methods use the straight-through estimator (STE) to train quantized networks, which avoids a zero-gradient problem by replacing a derivative of a discretizer (i.e., a round function) with that of an identity function. Although quantized networks exploiting the STE have shown decent performance, the STE is sub-optimal in that it simply propagates the same gradient without considering discretization errors between inputs and outputs of the discretizer. In this paper, we propose an element-wise gradient scaling (EWGS), a simple yet effective alternative to the STE, training a quantized network better than the STE in terms of stability and accuracy. Given a gradient of the discretizer output, EWGS adaptively scales up or down each gradient element, and uses the scaled gradient as the one for the discretizer input to train quantized networks via backpropagation. The scaling is performed depending on both the sign of each gradient element and an error between the continuous input and discrete output of the discretizer. We adjust a scaling factor adaptively using Hessian information of a network. We show extensive experimental results on the image classification datasets, including CIFAR-10 and ImageNet, with diverse network architectures under a wide range of bit-width settings, demonstrating the effectiveness of our method.
****
img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation
Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, Tal Hassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7617-7627
We propose real-time, six degrees of freedom (6DoF), 3D face pose estimation without face detection or landmark localization. We observe that estimating the 6DoF rigid transformation of a face is a simpler problem than facial landmark detection, often used for 3D face alignment. In addition, 6DoF offers more information than face bounding box labels. We leverage these observations to make multiple contributions: (a) We describe an easily trained, efficient, Faster R-CNN-based model which regresses 6DoF pose for all faces in the photo, without preliminary face detection. (b) We explain how pose is converted and kept consistent between the input photo and arbitrary crops created while training and evaluating our model. (c) Finally, we show how face poses can replace detection bounding box training labels. Tests on AFLW2000-3D and BIWI show that our method runs at real-time and outperforms state of the art (SotA) face pose estimators. Remarkably, our method also surpasses SotA models of comparable complexity on the WIDER FACE detection benchmark, despite not been optimized on bounding box labels.
****
Sparse Multi-Path Corrections in Fringe Projection Profilometry
Yu Zhang, Daniel Lau, David Wipf; Proceedings of the IEEE/CVF Conference on Comp

uter Vision and Pattern Recognition (CVPR), 2021, pp. 13344-13353
Three-dimensional scanning by means of structured light illumination is an active imaging technique involving projecting and capturing a series of striped patterns and then using the observed warping of stripes to reconstruct the target object's surface through triangulating each pixel in the camera to a unique projector coordinate corresponding to a particular feature in the projected patterns. The undesirable phenomenon of multi-path occurs when a camera pixel simultaneously sees features from multiple projector coordinates. Bimodal multi-path is a particularly common situation found along step edges, where the camera pixel sees both a foreground and background surface. Generalized from bimodal multi-path, this paper looks at sparse or N modal multi-path as a more general case, where the camera pixel sees no less than two reflective surfaces, resulting in decoding errors. Using fringe projection profilometry, our proposed solution is to treat each camera pixel as an underdetermined linear system of equations and to find the sparsest (least number of paths) solution using an application-specific Bayesian learning approach. We validate this algorithm with both simulations and a number of challenging real-world scenarios, outperforming the state-of-the-art techniques.
****
NeuroMorph: Unsupervised Shape Interpolation and Correspondence in One Go
Marvin Eisenberger, David Novotny, Gael Kerchenbaum, Patrick Labatut, Natalia Neverova, Daniel Cremers, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7473-7483
We present NeuroMorph, a new neural network architecture that takes as input two 3D shapes and produces in one go, i.e. in a single feed forward pass, a smooth interpolation and point-to-point correspondences between them. The interpolation, expressed as a deformation field, changes the pose of the source shape to resemble the target, but leaves the object identity unchanged. NeuroMorph uses an elegant architecture combining graph convolutions with global feature pooling to extract local features. During training, the model is incentivized to create realistic deformations by approximating geodesics on the underlying shape space manifold. This strong geometric prior allows to train our model end-to-end and in a fully unsupervised manner without requiring any manual correspondence annotations. NeuroMorph works well for a large variety of input shapes, including non-isometric pairs from different object categories. It obtains state-of-the-art results for both shape correspondence and interpolation tasks, matching or surpassing the performance of recent unsupervised and supervised methods on multiple benchmarks.
****
Soft-IntroVAE: Analyzing and Improving the Introspective Variational Autoencoder
Tal Daniel, Aviv Tamar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4391-4400
The recently introduced introspective variational autoencoder (IntroVAE) exhibits outstanding image generations, and allows for amortized inference using an image encoder. The main idea in IntroVAE is to train a VAE adversarially, using the VAE encoder to discriminate between generated and real data samples. However, the original IntroVAE loss function relied on a particular hinge-loss formulation that is very hard to stabilize in practice, and its theoretical convergence analysis ignored important terms in the loss. In this work, we take a step towards better understanding of the IntroVAE model, its practical implementation, and its applications. We propose the Soft-IntroVAE, a modified IntroVAE that replaces the hinge-loss terms with a smooth exponential loss on generated samples. This change significantly improves training stability, and also enables theoretical analysis of the complete algorithm. Interestingly, we show that the IntroVAE converges to a distribution that minimizes a sum of KL distance from the data distribution and an entropy term. We discuss the implications of this result, and demonstrate that it induces competitive image generation and reconstruction. Finally, we describe an application of Soft-IntroVAE to unsupervised image translation, and demonstrate compelling results. Code and additional information is available on the project website - taldatech.github.io/soft-intro-vae-web

****
Energy-Based Learning for Scene Graph Generation
Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, Leonid Sigal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13936-13945
Traditional scene graph generation methods are trained using cross-entropy losses that treat objects and relationships as independent entities. Such a formulation, however, ignores structure in the output space, in an inherently structured prediction problem. In this work, we introduce a novel energy-based learning framework for generating scene graphs. The proposed formulation allows for efficiently incorporating the structure of scene graphs in the output space. This additional constraint in the learning framework acts as an inductive bias and allows models to learn efficiently from a small number of labels. We use the proposed energy-based framework to train existing state-of-the-art models and show a significant performance improvement, of up to 21% and 27%, on the Visual Genome and GQA benchmark datasets, respectively. Further, we showcase the learning efficiency of the proposed framework by demonstrating superior performance in the zero- and few-shot settings where data is scarce.
****
Zillow Indoor Dataset: Annotated Floor Plans With 360deg Panoramas and 3D Room Layouts
Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, Sing Bing Kang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2133-2143
We present Zillow Indoor Dataset (ZInD): A large indoor dataset with 71,474 panoramas from 1,524 real unfurnished homes. ZInD provides annotations of 3D room layouts, 2D and 3D floor plans, panorama location in the floor plan, and locations of windows and doors. The ground truth construction took over 1,500 hours of annotation work. To the best of our knowledge, ZInD is the largest real dataset with layout annotations. A unique property is the room layout data, which follows a real world distribution (cuboid, more general Manhattan, and non-Manhattan layouts) as opposed to the mostly cuboid or Manhattan layouts in current publicly available datasets. Also, the scale and annotations provided are valuable for effective research related to room layout and floor plan analysis. To demonstrate ZInD's benefits, we benchmark on room layout estimation from single panoramas and multi-view registration.
****
Progressive Contour Regression for Arbitrary-Shape Scene Text Detection
Pengwen Dai, Sanyi Zhang, Hua Zhang, Xiaochun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7393-7402
State-of-the-art scene text detection methods usually model the text instance with local pixels or components from the bottom-up perspective and, therefore, are sensitive to noises and dependent on the complicated heuristic post-processing especially for arbitrary-shape texts. To relieve these two issues, instead, we propose to progressively evolve the initial text proposal to arbitrarily shaped text contours in a top-down manner. The initial horizontal text proposals are generated by estimating the center and size of texts. To reduce the range of regression, the first stage of the evolution predicts the corner points of oriented text proposals from the initial horizontal ones. In the second stage, the contours of the oriented text proposals are iteratively regressed to arbitrarily shaped ones. In the last iteration of this stage, we rescore the confidence of the final localized text by utilizing the cues from multiple contour points, rather than the single cue from the initial horizontal proposal center that may be out of arbitrary-shape text regions. Moreover, to facilitate the progressive contour evolution, we design a contour information aggregation mechanism to enrich the feature representation on text contours by considering both the circular topology and semantic context. Experiments conducted on CTW1500, Total-Text, ArT, and TD500 have demonstrated that the proposed method especially excels in line-level arbitrary-shape texts. Code is available at http://github.com/dpengwen/PCR.
****

UV-Net: Learning From Boundary Representations
Pradeep Kumar Jayaraman, Aditya Sanghi, Joseph G. Lambourne, Karl D.D. Willis, Thomas Davies, Hooman Shayani, Nigel Morris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11703-11712
We introduce UV-Net, a novel neural network architecture and representation designed to operate directly on Boundary representation (B-rep) data from 3D CAD models. The B-rep format is widely used in the design, simulation and manufacturing industries to enable sophisticated and precise CAD modeling operations. However, B-rep data presents some unique challenges when used with modern machine learning due to the complexity of the data structure and its support for both continuous non-Euclidean geometric entities and discrete topological entities. In this paper, we propose a unified representation for B-rep data that exploits the U and V parameter domain of curves and surfaces to model geometry, and an adjacency graph to explicitly model topology. This leads to a unique and efficient network architecture, UV-Net, that couples image and graph convolutional neural networks in a compute and memory-efficient manner. To aid in future research we present a synthetic labelled B-rep dataset, SolidLetters, derived from human designed fonts with variations in both geometry and topology. Finally we demonstrate that UV-Net can generalize to supervised and unsupervised tasks on five datasets, while outperforming alternate 3D shape representations such as point clouds, voxels, and meshes.
****
MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation
Sanjay Kariyappa, Atul Prakash, Moinuddin K Qureshi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13814-13823
High quality Machine Learning (ML) models are often considered valuable intellectual property by companies. Model Stealing (MS) attacks allow an adversary with black-box access to a ML model to replicate its functionality by training a clone model using the predictions of the target model for different inputs. However, best available existing MS attacks fail to produce a high-accuracy clone without access to the target dataset or a representative dataset necessary to query the target model. In this paper, we show that preventing access to the target dataset is not an adequate defense to protect a model. We propose MAZE -- a data-free model stealing attack using zeroth-order gradient estimation that produces high-accuracy clones. In contrast to prior works, MAZE uses only synthetic data created using a generative model to perform MS. Our evaluation with four image classification models shows that MAZE provides a normalized clone accuracy in the range of 0.90x to 0.99x, and outperforms even the recent attacks that rely on partial data (JBDA, clone accuracy 0.13x to 0.69x) and on surrogate data (KnockoffNets, clone accuracy 0.52x to 0.97x). We also study an extension of MAZE in the partial-data setting and develop MAZE-PD, which generates synthetic data closer to the target distribution. MAZE-PD further improves the clone accuracy 0.97x to 1.0x) and reduces the query budget required for the attack by 2x-24x.
****
Universal Spectral Adversarial Attacks for Deformable Shapes
Arianna Rampini, Franco Pestarini, Luca Cosmo, Simone Melzi, Emanuele Rodola; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3216-3226
Machine learning models are known to be vulnerable to adversarial attacks, namely perturbations of the data that lead to wrong predictions despite being imperceptible. However, the existence of "universal" attacks (i.e., unique perturbations that transfer across different data points) has only been demonstrated for images to date. Part of the reason lies in the lack of a common domain, for geometric data such as graphs, meshes, and point clouds, where a universal perturbation can be defined. In this paper, we offer a change in perspective and demonstrate the existence of universal attacks for geometric data (shapes). We introduce a computational procedure that operates entirely in the spectral domain, where the attacks take the form of small perturbations to short eigenvalue sequences; the resulting geometry is then synthesized via shape-from-spectrum recovery. Our at

tacks are universal, in that they transfer across different shapes, different representations (meshes and point clouds), and generalize to previously unseen data.
****
Prototypical Cross-Domain Self-Supervised Learning for Few-Shot Unsupervised Domain Adaptation
Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, Alberto Sangiovanni Vincentelli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13834-13844
Unsupervised Domain Adaptation (UDA) transfers predictive models from a fully-labeled source domain to an unlabeled target domain. In some applications, however, it is expensive even to collect labels in the source domain, making most previous works impractical. To cope with this problem, recent work performed instance-wise cross-domain self-supervised learning, followed by an additional fine-tuning stage. However, the instance-wise self-supervised learning only learns and aligns low-level discriminative features. In this paper, we propose an end-to-end Prototypical Cross-domain Self-Supervised Learning (PCS) framework for Few-shot Unsupervised Domain Adaptation (FUDA). PCS not only performs cross-domain low-level feature alignment, but it also encodes and aligns semantic structures in the shared embedding space across domains. Our framework captures category-wise semantic structures of the data by in-domain prototypical contrastive learning; and performs feature alignment through cross-domain prototypical self-supervision. Compared with state-of-the-art methods, PCS improves the mean classification accuracy over different domain pairs on FUDA by 10.5%, 3.5%, 9.0%, and 13.2% on Office, Office-Home, VisDA-2017, and DomainNet, respectively.
****
HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation
Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3383-3393
Model-based 3D pose and shape estimation methods reconstruct a full 3D mesh for the human body by estimating several parameters. However, learning the abstract parameters is a highly non-linear process and suffers from image-model misalignment, leading to mediocre model performance. In contrast, 3D keypoint estimation methods combine deep CNN network with the volumetric representation to achieve pixel-level localization accuracy but may predict unrealistic body structure. In this paper, we address the above issues by bridging the gap between body mesh estimation and 3D keypoint estimation. We propose a novel hybrid inverse kinematics solution (HybrIK). HybrIK directly transforms accurate 3D joints to relative body-part rotations for 3D body mesh reconstruction, via the twist-and-swing decomposition. The swing rotation is analytically solved with 3D joints, and the twist rotation is derived from the visual cues through the neural network. We show that HybrIK preserves both the accuracy of 3D pose and the realistic body structure of the parametric human model, leading to a pixel-aligned 3D body mesh and a more accurate 3D pose than the pure 3D keypoint estimation methods. Without bells and whistles, the proposed method surpasses the state-of-the-art methods by a large margin on various 3D human pose and shape benchmarks. As an illustrative example, HybrIK outperforms all the previous methods by 13.2 mm MPJPE and 21.9 mm PVE on 3DPW dataset. Our code is available at https://github.com/Jeff-sjtu/HybrIK.
****
Human De-Occlusion: Invisible Perception and Recovery for Humans
Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, Xinggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3691-3701
In this paper, we tackle the problem of human de-occlusion which reasons about occluded segmentation masks and invisible appearance content of humans. In particular, a two-stage framework is proposed to estimate the invisible portions and recover the content inside. For the stage of mask completion, a stacked network s

tructure is devised to refine inaccurate masks from a general instance segmentation model and predict integrated masks simultaneously. Additionally, the guidance from human parsing and typical pose masks are leveraged to bring prior information. For the stage of content recovery, a novel parsing guided attention module is applied to isolate body parts and capture context information across multiple scales. Besides, an Amodal Human Perception dataset (AHP) is collected to settle the task of human de-occlusion. AHP has advantages of providing annotations from real-world scenes and the number of humans is comparatively larger than other amodal perception datasets. Based on this dataset, experiments demonstrate that our method performs over the state-of-the-art techniques in both tasks of mask completion and content recovery. Our AHP dataset is available at https://sydney0zq.github.io/ahp/.
****

The Neural Tangent Link Between CNN Denoisers and Non-Local Filters
Julian Tachella, Junqi Tang, Mike Davies; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8618-8627
Convolutional Neural Networks (CNNs) are now a well-established tool for solving computational imaging problems. Modern CNN-based algorithms obtain state-of-the-art performance in diverse image restoration problems. Furthermore, it has been recently shown that, despite being highly overparameterized, networks trained with a single corrupted image can still perform as well as fully trained networks. We introduce a formal link between such networks through their neural tangent kernel (NTK), and well-known non-local filtering techniques, such as non-local means or BM3D. The filtering function associated with a given network architecture can be obtained in closed form without need to train the network, being fully characterized by the random initialization of the network weights. While the NTK theory accurately predicts the filter associated with networks trained using standard gradient descent, our analysis shows that it falls short to explain the behaviour of networks trained using the popular Adam optimizer. The latter achieves a larger change of weights in hidden layers, adapting the non-local filtering function during training. We evaluate our findings via extensive image denoising experiments.
****

Achieving Robustness in Classification Using Optimal Transport With Hinge Regularization
Mathieu Serrurier, Franck Mamalet, Alberto Gonzalez-Sanz, Thibaut Boissin, Jean-Michel Loubes, Eustasio del Barrio; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 505-514
Adversarial examples have pointed out Deep Neural Network's vulnerability to small local noise. It has been shown that constraining their Lipschitz constant should enhance robustness, but make them harder to learn with classical loss functions. We propose a new framework for binary classification, based on optimal transport, which integrates this Lipschitz constraint as a theoretical requirement. We propose to learn 1-Lipschitz networks using a new loss that is an hinge regularized version of the Kantorovich-Rubinstein dual formulation for the Wasserstein distance estimation. This loss function has a direct interpretation in terms of adversarial robustness together with certifiable robustness bound. We also prove that this hinge regularized version is still the dual formulation of an optimal transportation problem, and has a solution. We also establish several geometrical properties of this optimal solution, and extend the approach to multi-class problems. Experiments show that the proposed approach provides the expected guarantees in terms of robustness without any significant accuracy drop. The adversarial examples, on the proposed models, visibly and meaningfully change the input providing an explanation for the classification.
****

Stochastic Image-to-Video Synthesis Using cINNs
Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, Bjorn Ommer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3742-3753
Video understanding calls for a model to learn the characteristic interplay betw

een static scene content and its dynamics: Given an image, the model must be able to predict a future progression of the portrayed scene and, conversely, a video should be explained in terms of its static image content and all the remaining characteristics not present in the initial frame. This naturally suggests a bijective mapping between the video domain and the static content as well as residual information. In contrast to common stochastic image-to-video synthesis, such a model does not merely generate arbitrary videos progressing the initial image. Given this image, it rather provides a one-to-one mapping between the residual vectors and the video with stochastic outcomes when sampling. The approach is naturally implemented using a conditional invertible neural network (cINN) that can explain videos by independently modelling static and other video characteristics, thus laying the basis for controlled video synthesis. Experiments on diverse video datasets demonstrate the effectiveness of our approach in terms of both the quality and diversity of the synthesized results. Our project page is available at https://bit.ly/3dg90fV.
****
Ego-Exo: Transferring Visual Representations From Third-Person to First-Person Videos
Yanghao Li, Tushar Nagarajan, Bo Xiong, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6943-6953
We introduce an approach for pre-training egocentric video models using large-scale third-person video datasets. Learning from purely egocentric data is limited by low dataset scale and diversity, while using purely exocentric (third-person) data introduces a large domain mismatch. Our idea is to discover latent signals in third-person video that are predictive of key egocentric-specific properties. Incorporating these signals as knowledge distillation losses during pre-training results in models that benefit from both the scale and diversity of third-person video data, as well as representations that capture salient egocentric properties. Our experiments show that our Ego-Exo framework can be seamlessly integrated into standard video models; it outperforms all baselines when fine-tuned for egocentric activity recognition, achieving state-of-the-art results on Charades-Ego and EPIC-Kitchens-100.
****
Dynamic Slimmable Network
Changlin Li, Guangrun Wang, Bing Wang, Xiaodan Liang, Zhihui Li, Xiaojun Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8607-8617
Current dynamic networks and dynamic pruning methods have shown their promising capability in reducing theoretical computation complexity. However, dynamic sparse patterns on convolutional filters fail to achieve actual acceleration in real-world implementation, due to the extra burden of indexing, weight-copying, or zero-masking. Here, we explore a dynamic network slimming regime, named Dynamic Slimmable Network (DS-Net), which aims to achieve good hardware-efficiency via dynamically adjusting filter numbers of networks at test time with respect to different inputs, while keeping filters stored statically and contiguously in hardware to prevent the extra burden. Our DS-Net is empowered with the ability of dynamic inference by the proposed double-headed dynamic gate that comprises an attention head and a slimming head to predictively adjust network width with negligible extra computation cost. To ensure generality of each candidate architecture and the fairness of gate, we propose a disentangled two-stage training scheme inspired by one-shot NAS. In the first stage, a novel training technique for weight-sharing networks named In-place Ensemble Bootstrapping is proposed to improve the supernet training efficacy. In the second stage, Sandwich Gate Sparsification is proposed to assist the gate training by identifying easy and hard samples in an online way. Extensive experiments demonstrate our DS-Net consistently outperforms its static counterparts as well as state-of-the-art static and dynamic model compression methods by a large margin (up to 5.9%). Typically, DS-Net achieves 2-4x computation reduction and 1.62x real-world acceleration over ResNet-50 and MobileNet with minimal accuracy drops on ImageNet.

****
Jo-SRC: A Contrastive Approach for Combating Noisy Labels
Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, Zhenmin Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5192-5201
Due to the memorization effect in Deep Neural Networks (DNNs), training with noisy labels usually results in inferior model performance. Existing state-of-the-art methods primarily adopt a sample selection strategy, which selects small-loss samples for subsequent training. However, prior literature tends to perform sample selection within each mini-batch, neglecting the imbalance of noise ratios in different mini-batches. Moreover, valuable knowledge within high-loss samples is wasted. To this end, we propose a noise-robust approach named Jo-SRC (Joint Sample Selection and Model Regularization based on Consistency). Specifically, we train the network in a contrastive learning manner. Predictions from two different views of each sample are used to estimate its "likelihood" of being clean or out-of-distribution. Furthermore, we propose a joint loss to advance the model generalization performance by introducing consistency regularization. Extensive experiments and ablation studies have validated the superiority of our approach over existing state-of-the-art methods.
****
Deep Lucas-Kanade Homography for Multimodal Image Alignment
Yiming Zhao, Xinming Huang, Ziming Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15950-15959
Estimating homography to align image pairs captured by different sensors or image pairs with large appearance changes is an important and general challenge for many computer vision applications. In contrast to others, we propose a generic solution to pixel-wise align multimodal image pairs by extending the traditional Lucas-Kanade algorithm with networks. The key contribution in our method is how we construct feature maps, named as deep Lucas-Kanade feature map (DLKFM). The learned DLKFM can spontaneously recognize invariant features under various appearance-changing conditions. It also has two nice properties for the Lucas-Kanade algorithm: (1) The template feature map keeps brightness consistency with the input feature map, thus the color difference is very small while they are well-aligned. (2) The Lucas-Kanade objective function built on DLKFM has a smooth landscape around ground truth homography parameters, so the iterative solution of the Lucas-Kanade can easily converge to the ground truth. With those properties, directly updating the Lucas-Kanade algorithm on our feature maps will precisely align image pairs with large appearance changes. We share the dataset, code, and demo video online.
****
clDice - A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation
Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, Bjoern H. Menze; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16560-16569
Accurate segmentation of tubular, network-like structures, such as vessels, neurons, or roads, is relevant to many fields of research. For such structures, the topology is their most important characteristic; particularly preserving connectedness: in the case of vascular networks, missing a connected vessel entirely alters the blood-flow dynamics. We introduce a novel similarity measure termed centerlineDice (short clDice), which is calculated on the intersection of the segmentation masks and their (morphological) skeleta. We theoretically prove that clDice guarantees topology preservation up to homotopy equivalence for binary 2D and 3D segmentation. Extending this, we propose a computationally efficient, differentiable loss function (soft-clDice) for training arbitrary neural segmentation networks. We benchmark the soft-clDice loss on five public datasets, including vessels, roads and neurons (2D and 3D). Training on soft-clDice leads to segmentation with more accurate connectivity information, higher graph similarity, and better volumetric scores.

****
Hyper-LifelongGAN: Scalable Lifelong Learning for Image Conditioned Generation
Mengyao Zhai, Lei Chen, Greg Mori; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2246-2255
Deep neural networks are susceptible to catastrophic forgetting: when encountering a new task, they can only remember the new task and fail to preserve its ability to accomplish previously learned tasks. In this paper, we study the problem of lifelong learning for generative models and propose a novel and generic continual learning framework Hyper-LifelongGAN which is more scalable compared with state-of-the-art approaches. Given a sequence of tasks, the conventional convolutional filters are factorized into the dynamic base filters which are generated using task specific filter generators, and deterministic weight matrix which linearly combines the base filters and is shared across different tasks. Moreover, the shared weight matrix is multiplied by task specific coefficients to introduce more flexibility in combining task specific base filters differently for different tasks. Attributed to the novel architecture, the proposed method can preserve or even improve the generation quality at a low cost of parameters. We validate Hyper-LifelongGAN on diverse image-conditioned generation tasks, extensive ablation studies and comparisons with state-of-the-art models are carried out to show that the proposed approach can address catastrophic forgetting effectively.
****
Semi-Supervised Synthesis of High-Resolution Editable Textures for 3D Humans
Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, Tony Tung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7991-8000
We introduce a novel approach to generate diverse high fidelity texture maps for 3D human meshes in a semi-supervised setup. Given a segmentation mask defining the layout of the semantic regions in the texture map, our network generates high-resolution textures with a variety of styles, that are then used for rendering purposes. To accomplish this task, we propose a Region-adaptive Adversarial Variational AutoEncoder (ReAVAE) that learns the probability distribution of the style of each region individually so that the style of the generated texture can be controlled by sampling from the region-specific distributions. In addition, we introduce a data generation technique to augment our training set with data lifted from single-view RGB inputs. Our training strategy allows the mixing of reference image styles with arbitrary styles for different regions, a property which can be valuable for virtual try-on AR/VR applications. Experimental results show that our method synthesizes better texture maps compared to prior work while enabling independent layout and style controllability.
****
CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback
Seungmin Lee, Dongwan Kim, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 802-812
We tackle the task of image retrieval with text feedback, where a reference image and modifier text are combined to identify the desired target image. We focus on designing an image-text compositor, i.e., integrating multi-modal inputs to produce a representation similar to that of the target image. In our algorithm, Content-Style Modulation (CoSMo), we approach this challenge by introducing two modules based on deep neural networks: the content and style modulators. The content modulator performs local updates to the reference image feature after normalizing the style of the image, where a disentangled multi-modal non-local block is employed to achieve the desired content modifications. Then, the style modulator reintroduces global style information to the updated feature. We provide an in-depth view of our algorithm and its design choices, and show that it accomplishes outstanding performance on multiple image-text retrieval benchmarks. Our code can be found at: https://github.com/postBG/CosMo.pytorch
****
Thinking Fast and Slow: Efficient Text-to-Visual Retrieval With Transformers
Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit

ion (CVPR), 2021, pp. 9826-9836
Our objective is language-based search of large-scale image and video datasets. For this task, the approach that consists of independently mapping text and vision to a joint embedding space, a.k.a. dual encoders, is attractive as retrieval scales and is efficient for billions of images using approximate nearest neighbour search. An alternative approach of using vision-text transformers with cross-attention gives considerable improvements in accuracy over the joint embeddings, but is often inapplicable in practice for large-scale retrieval given the cost of the cross-attention mechanisms required for each sample at test time. This work combines the best of both worlds. We make the following three contributions. First, we equip transformer-based models with a new fine-grained cross-attention architecture, providing significant improvements in retrieval accuracy whilst preserving scalability. Second, we introduce a generic approach for combining a Fast dual encoder model with our Slow but accurate transformer-based model via distillation and re-ranking. Finally, we validate our approach on the Flickr30K image dataset where we show an increase in inference speed by several orders of magnitude while having results competitive to the state of the art. We also extend our method to the video domain, improving the state of the art on the VATEX dataset.
****
RGB-D Local Implicit Function for Depth Completion of Transparent Objects
Luyang Zhu, Arsalan Mousavian, Yu Xiang, Hammad Mazhar, Jozef van Eenbergen, Shoubhik Debnath, Dieter Fox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4649-4658
Majority of the perception methods in robotics require depth information provided by RGB-D cameras. However, standard 3D sensors fail to capture depth of transparent objects due to refraction and absorption of light. In this paper, we introduce a new approach for depth completion of transparent objects from a single RGB-D image. Key to our approach is a local implicit neural representation built on ray-voxel pairs that allows our method to generalize to unseen objects and achieve fast inference speed. Based on this representation, we present a novel framework that can complete missing depth given noisy RGB-D input. We further improve the depth estimation iteratively using a self-correcting refinement model. To train the whole pipeline, we build a large scale synthetic dataset with transparent objects. Experiments demonstrate that our method performs significantly better than the current state-of-the-art methods on both synthetic and real world data. In addition, our approach improves the inference speed by a factor of 20 compared to the previous best method, ClearGrasp. Code will be released at https://research.nvidia.com/publication/2021-03_RGB-D-Local-Implicit.
****
Fingerspelling Detection in American Sign Language
Bowen Shi, Diane Brentari, Greg Shakhnarovich, Karen Livescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4166-4175
Fingerspelling, in which words are signed letter by letter, is an important component of American Sign Language. Most previous work on automatic fingerspelling recognition has assumed that the boundaries of fingerspelling regions in signing videos are known beforehand. In this paper, we consider the task of fingerspelling detection in raw, untrimmed sign language videos. This is an important step towards building real-world fingerspelling recognition systems. We propose a benchmark and a suite of evaluation metrics, some of which reflect the effect of detection on the downstream fingerspelling recognition task. In addition, we propose a new model that learns to detect fingerspelling via multi-task training, incorporating pose estimation and fingerspelling recognition (transcription) along with detection, and compare this model to several alternatives. The model outperforms all alternative approaches across all metrics, establishing a state of the art on the benchmark.
****
Uncertainty Reduction for Model Adaptation in Semantic Segmentation
Prabhu Teja S, Francois Fleuret; Proceedings of the IEEE/CVF Conference on Compu

ter Vision and Pattern Recognition (CVPR), 2021, pp. 9613-9623
Traditional methods for Unsupervised Domain Adaptation (UDA) targeting semantic segmentation exploit information common to the source and target domains, using both labeled source data and unlabeled target data. In this paper, we investigate a setting where the source data is unavailable, but the classifier trained on the source data is; hence named ""model adaptation"". Such a scenario arises when data sharing is prohibited, for instance, because of privacy, or Intellectual Property (IP) issues. To tackle this problem, we propose a method that reduces the uncertainty of predictions on the target domain data. We accomplish this in two ways: minimizing the entropy of the predicted posterior, and maximizing the noise robustness of the feature representation. We show the efficacy of our method on the transfer of segmentation from computer generated images to real-world driving images, and transfer between data collected in different cities, and surprisingly reach performance competitive with that of the methods that have access to source data.
****
Learning Triadic Belief Dynamics in Nonverbal Communication From Videos
Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, Yixin Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7312-7321
Humans possess a unique social cognition capability; nonverbal communication can convey rich social information among agents. In contrast, such crucial social characteristics are mostly missing in the existing scene understanding literature. In this paper, we incorporate different nonverbal communication cues (e.g., gaze, human poses, and gestures) to represent, model, learn, and infer agents' mental states from pure visual inputs. Crucially, such a mental representation takes the agent's belief into account so that it represents what the true world state is and infers the beliefs in each agent's mental state, which may differ from the true world states. By aggregating different beliefs and true world states, our model essentially forms "five minds" during the interactions between two agents. This "five minds" model differs from prior works that infer beliefs in an infinite recursion; instead, agents' beliefs are converged into a "common mind". Based on this representation, we further devise a hierarchical energy-based model that jointly tracks and predicts all five minds. From this new perspective, a social event is interpreted by a series of nonverbal communication and belief dynamics, which transcends the classic keyframe video summary. In the experiments, we demonstrate that using such a social account provides a better video summary on videos with rich social interactions compared with state-of-the-art keyframe video summary methods.
****
Temporal Modulation Network for Controllable Space-Time Video Super-Resolution
Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6388-6397
Space-time video super-resolution (STVSR) aims to increase the spatial and temporal resolutions of low-resolution and low-frame-rate videos. Recently, deformable convolution based methods have achieved promising STVSR performance, but they could only infer the intermediate frame pre-defined in the training stage. Besides, these methods undervalued the short-term motion cues among adjacent frames. In this paper, we propose a Temporal Modulation Network (TMNet) to interpolate arbitrary intermediate frame(s) with accurate high-resolution reconstruction. Specifically, we propose a Temporal Modulation Block (TMB) to modulate deformable convolution kernels for controllable feature interpolation. To well exploit the temporal information, we propose a Locally-temporal Feature Comparison (LFC) module, along with the Bi-directional Deformable ConvLSTM, to extract short-term and long-term motion cues in videos. Experiments on three benchmark datasets demonstrate that our TMNet outperforms previous STVSR methods. The code is available at https://github.com/CS-GangXu/TMNet.
****
Zero-Shot Single Image Restoration Through Controlled Perturbation of Koschmiede

r's Model
Aupendu Kar, Sobhan Kanti Dhara, Debashis Sen, Prabir Kumar Biswas; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
21, pp. 16205-16215
Real-world image degradation due to light scattering can be described based on t
he Koschmieder's model. Training deep models to restore such degraded images is
challenging as real-world paired data is scarcely available and synthetic paired
 data may suffer from domain-shift issues. In this paper, a zero-shot single rea
l-world image restoration model is proposed leveraging a theoretically deduced p
roperty of degradation through the Koschmieder's model. Our zero-shot network es
timates the parameters of the Koschmieder's model, which describes the degradati
on in the input image, to perform image restoration. We show that a suitable deg
radation of the input image amounts to a controlled perturbation of the Koschmie
der's model that describes the image's formation. The optimization of the zero-s
hot network is achieved by seeking to maintain the relation between its estimate
s of Koschmieder's model parameters before and after the controlled perturbation
, along with the use of a few no-reference losses. Image dehazing and underwater
 image restoration are carried out using the proposed zero-shot framework, which
 in general outperforms the state-of-the-art quantitatively and subjectively on
multiple standard real-world image datasets. Additionally, the application of ou
r zero-shot framework for low-light image enhancement is also demonstrated.
****
Uncertainty-Aware Camera Pose Estimation From Points and Lines
Alexander Vakhitov, Luis Ferraz, Antonio Agudo, Francesc Moreno-Noguer; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2021, pp. 4659-4668
Perspective-n-Point-and-Line (PnPL) algorithms aim at fast, accurate, and robust
 camera localization with respect to a 3D model from 2D-3D feature correspondenc
es, being a major part of modern robotic and AR/VR systems. Current point-based
pose estimation methods use only 2D feature detection uncertainties, and the lin
e-based methods do not take uncertainties into account. In our setup, both 3D co
ordinates and 2D projections of the features are considered uncertain. We propos
e PnP(L) solvers based on EPnP[20] and DLS[14] for the uncertainty-aware pose es
timation. We also modify motion-only bundle adjustment to take 3D uncertainties
into account. We perform exhaustive synthetic and real experiments on two differ
ent visual odometry datasets. The new PnP(L) methods outperform the state-of-the
-art on real data in isolation, showing an increase in mean translation accuracy
 by 18% on a representative subset of KITTI, while the new uncertain refinement
improves pose accuracy for most of the solvers, e.g. decreasing mean translation
 error for the EPnP by 16% compared to the standard refinement on the same datas
et. The code is available at https://alexandervakhitov.github.io/uncertain-pnp/.
****
Temporal Context Aggregation Network for Temporal Action Proposal Refinement
Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao
, Junjie Yan, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 485-494
Temporal action proposal generation aims to estimate temporal intervals of actio
ns in untrimmed videos, which is a challenging yet important task in the video u
nderstanding field. The proposals generated by current methods still suffer from
 inaccurate temporal boundaries and inferior confidence used for retrieval owing
 to the lack of efficient temporal modeling and effective boundary context utili
zation. In this paper, we propose Temporal Context Aggregation Network (TCANet)
to generate high-quality action proposals through local and global temporal cont
ext aggregation and complementary as well as progressive boundary refinement. Sp
ecifically, we first design a Local-Global Temporal Encoder (LGTE), which adopts
 the channel grouping strategy to efficiently encode both local and global tempo
ral inter-dependencies. Furthermore, both the boundary and internal context of p
roposals are adopted for frame-level and segment-level boundary regressions, res
pectively. Temporal Boundary Regressor (TBR) is designed to combine these two re
gression granularities in an end-to-end fashion, which achieves the precise boun

daries and reliable confidence of proposals through progressive refinement. Extensive experiments are conducted on three challenging datasets: HACS, ActivityNet-v1.3, and THUMOS-14, where TCANet can generate proposals with high precision and recall. By combining with the existing action classifier, TCANet can obtain remarkable temporal action detection performance compared with other methods. Not surprisingly, the proposed TCANet won the 1st place in the CVPR 2020 - HACS challenge leaderboard on temporal action localization task.
****

Information-Theoretic Segmentation by Inpainting Error Maximization
Pedro Savarese, Sunnie S. Y. Kim, Michael Maire, Greg Shakhnarovich, David McAllester; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4029-4039
We study image segmentation from an information-theoretic perspective, proposing a novel adversarial method that performs unsupervised segmentation by partitioning images into maximally independent sets. More specifically, we group image pixels into foreground and background, with the goal of minimizing predictability of one set from the other. An easily computed loss drives a greedy search process to maximize inpainting error over these partitions. Our method does not involve training deep networks, is computationally cheap, class-agnostic, and even applicable in isolation to a single unlabeled image. Experiments demonstrate that it achieves a new state-of-the-art in unsupervised segmentation quality, while being substantially faster and more general than competing approaches.
****

Adaptive Prototype Learning and Allocation for Few-Shot Segmentation
Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, Joongkyu Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8334-8343
Prototype learning is extensively used for few-shot segmentation. Typically, a single prototype is obtained from the support feature by averaging the global object information. However, using one prototype to represent all the information may lead to ambiguities. In this paper, we propose two novel modules, named superpixel-guided clustering (SGC) and guided prototype allocation (GPA), for multiple prototype extraction and allocation. Specifically, SGC is a parameter-free and training-free approach, which extracts more representative prototypes by aggregating similar feature vectors, while GPA is able to select matched prototypes to provide more accurate guidance. By integrating the SGC and GPA together, we propose the Adaptive Superpixel-guided Network (ASGNet), which is a lightweight model and adapts to object scale and shape variation. In addition, our network can easily generalize to k-shot segmentation with substantial improvement and no additional computational cost. In particular, our evaluations on COCO demonstrate that ASGNet surpasses the state-of-the-art method by 5% in 5-shot segmentation.
****

RefineMask: Towards High-Quality Instance Segmentation With Fine-Grained Features
Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, Xiaolin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6861-6869
The two-stage methods for instance segmentation, e.g. Mask R-CNN, have achieved excellent performance recently. However, the segmented masks are still very coarse due to the downsampling operations in both the feature pyramid and the instance-wise pooling process, especially for large objects. In this work, we propose a new method called RefineMask for high-quality instance segmentation of objects and scenes, which incorporates fine-grained features during the instance-wise segmenting process in a multi-stage manner. Through fusing more detailed information stage by stage, RefineMask is able to refine high-quality masks consistently. RefineMask succeeds in segmenting hard cases such as bent parts of objects that are over-smoothed by most previous methods and outputs accurate boundaries. Without bells and whistles, RefineMask yields significant gains of 2.6, 3.4, 3.8 AP over Mask R-CNN on COCO, LVIS, and Cityscapes benchmarks respectively at a small amount of additional computational cost. Furthermore, our single-model result

outperforms the winner of the LVIS Challenge 2020 by 1.3 points on the LVIS test-dev set and establishes a new state-of-the-art. Code will be available at https://github.com/zhanggang001/RefineMask.
****
DCNAS: Densely Connected Neural Architecture Search for Semantic Image Segmentation
Xiong Zhang, Hongmin Xu, Hong Mo, Jianchao Tan, Cheng Yang, Lei Wang, Wenqi Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13956-13967
Existing NAS methods for dense image prediction tasks usually compromise on restricted search space or search on proxy task to meet the achievable computational demands. To allow as wide as possible network architectures and avoid the gap between realistic and proxy setting, we propose a novel Densely Connected NAS (DCNAS) framework, which directly searches the optimal network structures for the multi-scale representations of visual information, over a large-scale target dataset without proxy. Specifically, by connecting cells with each other using learnable weights, we introduce a densely connected search space to cover an abundance of mainstream network designs. Moreover, by combining both path-level and channel-level sampling strategies, we design a fusion module and mixture layer to reduce the memory consumption of ample search space, hence favoring the proxyless searching. Compared with contemporary works, experiments reveal that the proxyless searching scheme is capable of bridging the gap between searching and training environments. Further, DCNAS achieves new state-of-the-art performances on public semantic image segmentation benchmarks, including 84.3% on Cityscapes, and 86.9% on PASCAL VOC 2012. We also retain leading performances when evaluating the architecture on the more challenging ADE20K and PASCAL-Context dataset.
****
Tackling the Ill-Posedness of Super-Resolution Through Adaptive Target Generation
Younghyun Jo, Seoung Wug Oh, Peter Vajda, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16236-16245
By the one-to-many nature of the super-resolution (SR) problem, a single low-resolution (LR) image can be mapped to many high-resolution (HR) images. However, learning based SR algorithms are trained to map an LR image to the corresponding ground truth (GT) HR image in the training dataset. The training loss will increase and penalize the algorithm when the output does not exactly match the GT target, even when the outputs are mathematically valid candidates according to the SR framework. This becomes more problematic for the blind SR, as diverse unknown blur kernels exacerbate the ill-posedness of the problem. To this end, we propose a fundamentally different approach for the SR by introducing the concept of the adaptive target. The adaptive target is generated from the original GT target by a transformation to match the output of the SR network. The adaptive target provides an effective way for the SR algorithm to deal with the ill-posed nature of the SR, by providing the algorithm with the flexibility of accepting a variety of valid solutions. Experimental results show the effectiveness of our algorithm, especially for improving the perceptual quality of HR outputs.
****
DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation
Yufan He, Dong Yang, Holger Roth, Can Zhao, Daguang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5841-5850
Recently, neural architecture search(NAS) has been applied to automatically search high-performance networks for medical image segmentation. The NAS search space usually contains a network topology level(controlling connections among cells with different spatial scales) and a cell level(operations within each cell). Existing methods either require long searching time for large-scale 3D image datasets, or are limited to pre-defined topologies (such as U-shaped or single-path). In this work, we focus on three important aspects of NAS in 3D medical image se

gmentation: flexible multi-path network topology, high search efficiency, and bu
dgeted GPU memory usage. A novel differentiable search framework is proposed to
support fast gradient-based search within a highly flexible network topology sea
rch space. The discretization of the searched optimal continuous model in differ
entiable scheme may produce a sub-optimal final discrete model (discretization g
ap). Therefore, we propose a topology loss to alleviate this problem. In additio
n, the GPU memory usage for the searched 3D model is limited with budget constra
ints during search. Our Differentiable Network Topology Search scheme(DiNTS) is
evaluated on the Medical Segmentation Decathlon (MSD) challenge, which contains
ten challenging segmentation tasks. Our method achieves the state-of-the-art per
formance and the top ranking on the MSD challenge leaderboard.
****
Im2Vec: Synthesizing Vector Graphics Without Vector Supervision
Pradyumna Reddy, Michael Gharbi, Michal Lukac, Niloy J. Mitra; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p
p. 7342-7351
Vector graphics are widely used to represent fonts, logos, digital artworks, and
 graphic designs. But, while a vast body of work has focused on generative algor
ithms for raster images, only a handful of options exists for vector graphics. O
ne can always rasterize the input graphic and resort to image-based generative a
pproaches, but this negates the advantages of the vector representation. The cur
rent alternative is to use specialized models that require explicit supervision
on the vector graphics representation at training time. This is not ideal becaus
e large-scale high-quality vector-graphics datasets are difficult to obtain. Fur
thermore, the vector representation for a given design is not unique, so models
that supervise on the vector representation are unnecessarily constrained. Inste
ad, we propose a new neural network that can generate complex vector graphics wi
th varying topologies, and only requires in-direct supervision from readily-avai
lable raster training images (i.e., with no vector counterparts). To enable this
, we use a differentiable rasterization pipeline that renders the generated vect
or shapes and composites them together onto a raster canvas. We demonstrate our
method on a range of datasets, and provide comparison with state-of-the-art SVG-
VAE and DeepSVG, both of which require explicit vector graphics supervision. Fin
ally, we also demonstrate our approach on the MNIST dataset, for which no ground
truth vector representation is available.
****
Perception Matters: Detecting Perception Failures of VQA Models Using Metamorphi
c Testing
Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, Tsong Yueh Chen; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1
6908-16917
Visual question answering (VQA) takes an image and a natural-language question a
s input and returns a natural-language answer. To date, VQA models are primarily
 assessed by their accuracy on high-level reasoning questions. Nevertheless, Giv
en that perception tasks (e.g., recognizing objects) are the building blocks in
the compositional process required by high-level reasoning, there is a demanding
 need to gain insights into how much of a problem low-level perception is. Inspi
red by the principles of software metamorphic testing, we introduce MetaVQA, a m
odel-agnostic framework for benchmarking perception capability of VQA models. Gi
ven an image i, MetaVQA is able to synthesize a low level perception question q.
 It then jointly transforms (i, q) to one or a set of sub-questions and sub-imag
es. MetaVQA checks whether the answer to (i, q) satisfies metamorphic relationsh
ips (MRs), denoting perception consistency, with the composed answers of transfo
rmed questions and images. Violating MRs denotes a failure of answering percepti
on questions. MetaVQA successfully detects over 4.9 million perception failures
made by popular VQA models with metamorphic testing. The state-of-the-art VQA mo
dels (e.g., the champion of VQA 2020 Challenge) suffer from perception consisten
cy problems. In contrast, the Oscar VQA models, by using anchor points to align
questions and images, show generally better consistency in perception tasks. We
hope MetaVQA will revitalize interest in enhancing the low-level perceptual abil

ities of VQA models, a cornerstone of high-level reasoning.
****
Unsupervised Part Segmentation Through Disentangling Appearance and Shape
Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8355-8364
We study the problem of unsupervised discovery and segmentation of object parts, which, as an intermediate local representation, are capable of finding intrinsic object structure and providing more explainable recognition results. Recent unsupervised methods have greatly relaxed the dependency on annotated data which are costly to obtain, but still rely on additional information such as object segmentation masks or saliency map. To remove such a dependency and further improve the part segmentation performance, we develop a novel approach by disentangling the appearance and shape representations of object parts followed with reconstruction losses without using additional object mask information. To avoid degenerated solutions, a bottleneck block is designed to squeeze and expand the appearance representation, leading to a more effective disentanglement between geometry and appearance. Combined with a self-supervised part classification loss and an improved geometry concentration constraint, we can segment more consistent parts with semantic meanings. Comprehensive experiments on a wide variety of objects such as face, bird, and PASCAL VOC objects demonstrate the effectiveness of the proposed method.
****
Adversarial Imaging Pipelines
Buu Phan, Fahim Mannan, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16051-16061
Adversarial attacks play a critical role in understanding deep neural network predictions and improving their robustness. Existing attack methods aim to deceive convolutional neural network (CNN)-based classifiers by manipulating RGB images that are fed directly to the classifiers. However, these approaches typically neglect the influence of the camera optics and image processing pipeline (ISP) that produce the network inputs. ISPs transform RAW measurements to RGB images and traditionally are assumed to preserve adversarial patterns. In fact, these low-level pipelines can destroy, introduce or amplify adversarial patterns that can deceive a downstream detector. As a result, optimized patterns can become adversarial for the classifier after being transformed by a certain camera ISP or optical lens system but not for others. In this work, we examine and develop such an attack that deceives a specific camera ISP while leaving others intact, using the same downstream classifier. We frame this camera-specific attack as a multi-task optimization problem, relying on a differentiable approximation for the ISP itself. We validate the proposed method using recent state-of-the-art automotive hardware ISPs, achieving 92% fooling rate when attacking a specific ISP. We demonstrate physical optics attacks with 90% fooling rate for a specific camera lens.
****
Adaptive Consistency Regularization for Semi-Supervised Transfer Learning
Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, Dejing Dou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6923-6932
While recent studies on semi-supervised learning have shown remarkable progress in leveraging both labeled and unlabeled data, most of them presume a basic setting of the model is randomly initialized. In this work, we consider semi-supervised learning and transfer learning jointly, leading to a more practical and competitive paradigm that can utilize both powerful pre-trained models from the source domain as well as labeled/unlabeled data in the target domain. To better exploit the value of both pre-trained weights and unlabeled target examples, we introduce adaptive consistency regularization that consists of two complementary components: Adaptive Knowledge Consistency (AKC) on the examples between the source and target model, and Adaptive Representation Consistency (ARC) on the target model between labeled and unlabeled examples. Examples involved in the consistenc

y regularization are adaptively selected according to their potential contributions to the target task. We conduct extensive experiments on popular benchmarks including CIFAR-10, CUB-200, and MURA, by fine-tuning the ImageNet pre-trained ResNet-50 model. Results show that our proposed adaptive consistency regularization outperforms state-of-the-art semi-supervised learning techniques such as Pseudo Label, Mean Teacher, and FixMatch. Moreover, our algorithm is orthogonal to existing methods and thus able to gain additional improvements on top of MixMatch and FixMatch. Our code is available at https://github.com/Walleclipse/Semi-Supervised-Transfer-Learning-Paddle.
****
GANmut: Learning Interpretable Conditional Space for Gamut of Emotions
Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 568-577
Humans can communicate emotions through a plethora of facial expressions, each with its own intensity, nuances and ambiguities. The generation of such variety by means of conditional GANs is limited to the expressions encoded in the used label system. These limitations are caused either due to burdensome labeling demand or the confounded label space. On the other hand, learning from inexpensive and intuitive basic categorical emotion labels leads to limited emotion variability. In this paper, we propose a novel GAN-based framework which learns an expressive and interpretable conditional space (usable as a label space) of emotions, instead of conditioning on handcrafted labels. Our framework only uses the categorical labels of basic emotions to jointly learn the conditional space as well as the emotion manipulation. Such learning can benefit from the image variability within discrete labels, especially when the intrinsic labels reside beyond the discrete space of the defined. Our experiments demonstrate the effectiveness of the proposed framework, by allowing us to control and generate a gamut of complex and compound emotions, while using only the basic categorical emotion labels during training.
****
StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation
Zongze Wu, Dani Lischinski, Eli Shechtman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12863-12872
We explore and analyze the latent style space of StyleGAN2, a state-of-the-art architecture for image generation, using models pretrained on several different datasets. We first show that StyleSpace, the space of channel-wise style parameters, is significantly more disentangled than the other intermediate latent spaces explored by previous works. Next, we describe a method for discovering a large collection of style channels, each of which is shown to control a distinct visual attribute in a highly localized and disentangled manner. Third, we propose a simple method for identifying style channels that control a specific attribute, using a pretrained classifier or a small number of example images. Manipulation of visual attributes via these StyleSpace controls is shown to be better disentangled than via those proposed in previous works. To show this, we make use of a newly proposed Attribute Dependency metric. Finally, we demonstrate the applicability of StyleSpace controls to the manipulation of real images. Our findings pave the way to semantically meaningful and well-disentangled image manipulations via simple and intuitive interfaces.
****
Rethinking the Heatmap Regression for Bottom-Up Human Pose Estimation
Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, Erjin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13264-13273
Heatmap regression has become the most prevalent choice for nowadays human pose estimation methods. The ground-truth heatmaps are usually constructed by covering all skeletal keypoints by 2D gaussian kernels. The standard deviations of these kernels are fixed. However, for bottom-up methods, which need to handle a large variance of human scales and labeling ambiguities, the current practice seems unreasonable. To better cope with these problems, we propose the scale-adaptive

heatmap regression (SAHR) method, which can adaptively adjust the standard deviation for each keypoint. In this way, SAHR is more tolerant of various human scales and labeling ambiguities. However, SAHR may aggravate the imbalance between fore-background samples, which potentially hurts the improvement of SAHR. Thus, we further introduce the weight-adaptive heatmap regression (WAHR) to help balance the fore-background samples. Extensive experiments show that SAHR together with WAHR largely improves the accuracy of bottom-up human pose estimation. As a result, we finally outperform the state-of-the-art model by +1.5AP and achieve 72.0 AP on COCO test-dev2017, which is comparable with the performances of most top-down methods. Source codes are available at https://github.com/greatlog/SWAHR-HumanPose.
****
From Semantic Categories to Fixations: A Novel Weakly-Supervised Visual-Auditory Saliency Detection Approach
Guotao Wang, Chenglizhao Chen, Deng-Ping Fan, Aimin Hao, Hong Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15119-15128
Thanks to the rapid advances in the deep learning techniques and the wide availability of large-scale training sets, the performances of video saliency detection models have been improving steadily and significantly. However, the deep learning based visual-audio fixation prediction is still in its infancy. At present, only a few visual-audio sequences have been furnished with real fixations being recorded in the real visual-audio environment. Hence, it would be neither efficiency nor necessary to re-collect real fixations under the same visual-audio circumstance. To address the problem, this paper advocate a novel approach in a weakly-supervised manner to alleviating the demand of large-scale training sets for visual-audio model training. By using the video category tags only, we propose the selective class activation mapping (SCAM), which follows a coarse-to-fine strategy to select the most discriminative regions in the spatial-temporal-audio circumstance. Moreover, these regions exhibit high consistency with the real human-eye fixations, which could subsequently be employed as the pseudo GTs to train a new spatial-temporal-audio (STA) network. Without resorting to any real fixation, the performance of our STA network is comparable to that of the fully supervised ones.
****
High-Fidelity Face Tracking for AR/VR via Deep Lighting Adaptation
Lele Chen, Chen Cao, Fernando De la Torre, Jason Saragih, Chenliang Xu, Yaser Sheikh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13059-13069
3D video avatars can empower virtual communications by providing compression, privacy, entertainment, and a sense of presence in AR/VR. Best 3D photo-realistic AR/VR avatars driven by video, that can minimize uncanny effects, rely on person-specific models. However, existing person-specific photo-realistic 3D models are not robust to lighting, hence their results typically miss subtle facial behaviors and cause artifacts in the avatar. This is a major drawback for the scalability of these models in communication systems (e.g., Messenger, Skype, FaceTime) and AR/VR. This paper addresses previous limitations by learning a deep learning lighting model, that in combination with a high-quality 3D face tracking algorithm, provides a method for subtle and robust facial motion transfer from a regular video to a 3D photo-realistic avatar. Extensive experimental validation and comparisons to other state-of-the-art methods demonstrate the effectiveness of the proposed framework in real-world scenarios with variability in pose, expression, and illumination. Our project page can be found at https://www.cs.rochester.edu/ cxu22/r/wild-avatar/.
****
Mixed-Privacy Forgetting in Deep Networks
Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 792-801
We show that the influence of a subset of the training samples can be removed --

 or "forgotten" -- from the weights of a network trained on large-scale image cl
assification tasks, and we provide strong computable bounds on the amount of rem
aining information after forgetting. Inspired by real-world applications of forg
etting techniques, we introduce a novel notion of forgetting in mixed-privacy se
tting, where we know that a "core" subset of the training samples does not need
to be forgotten. While this variation of the problem is conceptually simple, we
show that working in this setting significantly improves the accuracy and guaran
tees of forgetting methods applied to vision classification tasks. Moreover, our
 method allows efficient removal of all information contained in non-core data b
y simply setting to zero a subset of the weights with minimal loss in performanc
e. We achieve these results by replacing a standard deep network with a suitable
 linear approximation. With opportune changes to the network architecture and tr
aining procedure, we show that such linear approximation achieves comparable per
formance to the original network and that the forgetting problem becomes quadrat
ic and can be solved efficiently even for large models. Unlike previous forgetti
ng methods on deep networks, ours can achieve close to the state-of-the-art accu
racy on large scale vision tasks. In particular, we show that our method allows
forgetting without having to trade off the model accuracy.
****
TediGAN: Text-Guided Diverse Face Image Generation and Manipulation
Weihao Xia, Yujiu Yang, Jing-Hao Xue, Baoyuan Wu; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2256-2265
In this work, we propose TediGAN, a novel framework for multi-modal image genera
tion and manipulation with textual descriptions. The proposed method consists of
 three components: StyleGAN inversion module, visual-linguistic similarity learn
ing, and instance-level optimization. The inversion module maps real images to t
he latent space of a well-trained StyleGAN. The visual-linguistic similarity lea
rns the text-image matching by mapping the image and text into a common embeddin
g space. The instance-level optimization is for identity preservation in manipul
ation. Our model can produce diverse and high-quality images with an unprecedent
ed resolution at 1024 x 1024. Using a control mechanism based on style-mixing, o
ur TediGAN inherently supports image synthesis with multi-modal inputs, such as
sketches or semantic labels, with or without instance guidance. To facilitate te
xt-guided multi-modal synthesis, we propose the Multi-Modal CelebA-HQ, a large-s
cale dataset consisting of real face images and corresponding semantic segmentat
ion map, sketch, and textual descriptions. Extensive experiments on the introduc
ed dataset demonstrate the superior performance of our proposed method. Code and
 data are available at https://github.com/weihaox/TediGAN.
****
Affective Processes: Stochastic Modelling of Temporal Context for Emotion and Fa
cial Expression Recognition
Enrique Sanchez, Mani Kumar Tellamekala, Michel Valstar, Georgios Tzimiropoulos;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2021, pp. 9074-9084
Temporal context is key to the recognition of expressions of emotion. Existing m
ethods, that rely on recurrent or self-attention models to enforce temporal cons
istency, work on the feature level, ignoring the task-specific temporal dependen
cies, and fail to model context uncertainty. To alleviate these issues, we build
 upon the framework of Neural Processes to propose a method for apparent emotion
 recognition with three key novel components: (a) probabilistic contextual repre
sentation with a global latent variable model; (b) temporal context modelling us
ing task-specific predictions in addition to features; and (c) smart temporal co
ntext selection. We validate our approach on four databases, two for Valence and
 Arousal estimation (SEWA and AffWild2), and two for Action Unit intensity estim
ation (DISFA and BP4D). Results show a consistent improvement over a series of s
trong baselines as well as over state-of-the-art methods.
****
ID-Unet: Iterative Soft and Hard Deformation for View Synthesis
Mingyu Yin, Li Sun, Qingli Li; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2021, pp. 7220-7229

View synthesis is usually done by an autoencoder, in which the encoder maps a source view image into a latent content code, and the decoder transforms it into a target view image according to the condition. However, the source contents are often not well kept in this setting, which leads to unnecessary changes during the view translation. Although adding skipped connections, like Unet, alleviates the problem, but it often causes the failure on the view conformity. This paper proposes a new architecture by performing the source-to-target deformation in an iterative way. Instead of simply incorporating the features from multiple layers of the encoder, we design soft and hard deformation modules, which warp the encoder features to the target view at different resolutions, and give results to the decoder to complement the details. Particularly, the current warping flow is not only used to align the feature of the same resolution, but also as an approximation to coarsely deform the high resolution feature. Then the residual flow is estimated and applied in the high resolution, so that the deformation is built up in the coarse-to-fine fashion. To better constrain the model, we synthesize a rough target view image based on the intermediate flows and their warped features. The extensive ablation studies and the final results on two different datasets show the effectiveness of the proposed model.
****
Positional Encoding As Spatial Inductive Bias in GANs
Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13569-13578
SinGAN shows impressive capability in learning internal patch distribution despite its limited effective receptive field. We are interested in knowing how such a translation-invariant convolutional generator could capture the global structure with just a spatially i.i.d. input. In this work, taking SinGAN and StyleGAN2 as examples, we show that such capability, to a large extent, is brought by the implicit positional encoding when using zero padding in the generators. Such positional encoding is indispensable for generating images with high fidelity. The same phenomenon is observed in other generative architectures such as DCGAN and PGGAN. We further show that zero padding leads to an unbalanced spatial bias with a vague relation between locations. To offer a better spatial inductive bias, we investigate alternative positional encodings and analyze their effects. Based on a more flexible positional encoding explicitly, we propose a new multi-scale training strategy and demonstrate its effectiveness in the state-of-the-art unconditional generator StyleGAN2. Besides, the explicit spatial inductive bias substantially improve SinGAN for more versatile image manipulation.
****
Mask-ToF: Learning Microlens Masks for Flying Pixel Correction in Time-of-Flight Imaging
Ilya Chugunov, Seung-Hwan Baek, Qiang Fu, Wolfgang Heidrich, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9116-9126
We introduce Mask-ToF, a method to reduce flying pixels (FP) in time-of-flight (ToF) depth captures. FPs are pervasive artifacts which occur around depth edges, where light paths from both an object and its background are integrated over the aperture. This light mixes at a sensor pixel to produce erroneous depth estimates, which can adversely affect downstream 3D vision tasks. Mask-ToF starts at the source of these FPs, learning a microlens-level occlusion mask which effectively creates a custom-shaped sub-aperture for each sensor pixel. This modulates the selection of foreground and background light mixtures on a per-pixel basis and thereby encodes scene geometric information directly into the ToF measurements. We develop a differentiable ToF simulator to jointly train a convolutional neural network to decode this information and produce high-fidelity, low-FP depth reconstructions. We test the effectiveness of Mask-ToF on a simulated light field dataset and validate the method with an experimental prototype. To this end, we manufacture the learned amplitude mask and design an optical relay system to virtually place it on a high-resolution ToF sensor. We find that Mask-ToF generalizes well to real data without retraining, cutting FP counts in half.

****
QPP: Real-Time Quantization Parameter Prediction for Deep Neural Networks
Vladimir Kryzhanovskiy, Gleb Balitskiy, Nikolay Kozyrskiy, Aleksandr Zuruev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10684-10692
Modern deep neural networks (DNNs) cannot be effectively used in mobile and embedded devices due to strict requirements for computational complexity, memory, and power consumption. The quantization of weights and feature maps (activations) is a popular approach to solve this problem. Training-aware quantization often shows excellent results but requires a full dataset, which is not always available. Post-training quantization methods, in turn, are applied without fine-tuning but still work well for many classes of tasks like classification, segmentation, and so on. However, they either imply a big overhead for quantization parameters (QPs) calculation at runtime (dynamic methods) or lead to an accuracy drop if pre-computed static QPs are used (static methods). Moreover, most inference frameworks don't support dynamic quantization. Thus we propose a novel quantization approach called QPP: quantization parameter prediction. With a small subset of a training dataset or unlabeled data from the same domain, we find the predictor that can accurately estimate QPs of activations given only the NN's input data. Such a predictor allows us to avoid complex calculation of precise values of QPs while maintaining the quality of the model. To illustrate our method's efficiency, we added QPP into two dynamic approaches: 1) Dense+Sparse quantization, where the predetermined percentage of activations are not quantized, 2) standard quantization with equal quantization steps. We provide experiments on a wide set of tasks including super-resolution, facial landmark, segmentation, and classification.
****
Nighttime Visibility Enhancement by Increasing the Dynamic Range and Suppression of Light Effects
Aashish Sharma, Robby T. Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11977-11986
Most existing nighttime visibility enhancement methods focus on low light. Night images, however, do not only suffer from low light, but also from man-made light effects such as glow, glare, floodlight, etc. Hence, when the existing nighttime visibility enhancement methods are applied to these images, they intensify the effects, degrading the visibility even further. High dynamic range (HDR) imaging methods can address the low light and over-exposed regions, however they cannot remove the light effects, and thus cannot enhance the visibility in the affected regions. In this paper, given a single nighttime image as input, our goal is to enhance its visibility by increasing the dynamic range of the intensity, and thus can boost the intensity of the low light regions, and at the same time, suppress the light effects (glow, glare) simultaneously. First, we use a network to estimate the camera response function (CRF) from the input image to linearise the image. Second, we decompose the linearised image into low-frequency (LF) and high-frequency (HF) feature maps that are processed separately through two networks for light effects suppression and noise removal respectively. Third, we use a network to increase the dynamic range of the processed LF feature maps, which are then combined with the processed HF feature maps to generate the final output that has increased dynamic range and suppressed light effects. Our experiments show the effectiveness of our method in comparison with the state-of-the-art nighttime visibility enhancement methods.
****
Self-Supervised Augmentation Consistency for Adapting Semantic Segmentation
Nikita Araslanov, Stefan Roth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15384-15394
We propose an approach to domain adaptation for semantic segmentation that is both practical and highly accurate. In contrast to previous work, we abandon the use of computationally involved adversarial objectives, network ensembles and style transfer. Instead, we employ standard data augmentation techniques - photometric noise, flipping and scaling - and ensure consistency of the semantic predict

ions across these image transformations. We develop this principle in a lightweight self-supervised framework trained on co-evolving pseudo labels without the need for cumbersome extra training rounds. Simple in training from a practitioner's standpoint, our approach is remarkably effective. We achieve significant improvements of the state-of-the-art segmentation accuracy after adaptation, consistent both across different choices of the backbone architecture and adaptation scenarios.
****
Patch-VQ: 'Patching Up' the Video Quality Problem
Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, Alan Bovik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14019-14029
No-reference (NR) perceptual video quality assessment (VQA) is a complex, unsolved, and important problem for social and streaming media applications. Efficient and accurate video quality predictors are needed to monitor and guide the processing of billions of shared, often imperfect, user-generated content (UGC). Unfortunately, current NR models are limited in their prediction capabilities on real-world, "in-the-wild" UGC video data. To advance progress on this problem, we created the largest (by far) subjective video quality dataset, containing 38,811 real-world distorted videos and 116,433 space-time localized video patches ('v-patches'), and 5.5M human perceptual quality annotations. Using this, we created two unique NR-VQA models: (a) a local-to-global region-based NR VQA architecture (called PVQ) that learns to predict global video quality and achieves state-of-the-art performance on 3 UGC datasets, and (b) a first-of-a-kind space-time video quality mapping engine (called PVQ Mapper) that helps localize and visualize perceptual distortions in space and time. The entire dataset and prediction models are freely available at https://live.ece.utexas.edu/research.php.
****
Double Low-Rank Representation With Projection Distance Penalty for Clustering
Zhiqiang Fu, Yao Zhao, Dongxia Chang, Xingxing Zhang, Yiming Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5320-5329
This paper presents a novel, simple yet robust self-representation method, i.e., Double Low-Rank Representation with Projection Distance penalty (DLRRPD) for clustering. With the learned optimal projected representations, DLRRPD is capable of obtaining an effective similarity graph to capture the multi-subspace structure. Besides the global low-rank constraint, the local geometrical structure is additionally exploited via a projection distance penalty in our DLRRPD, thus facilitating a more favorable graph. Moreover, to improve the robustness of DLRRPD to noises, we introduce a Laplacian rank constraint, which can further encourage the learned graph to be more discriminative for clustering tasks. Meanwhile, Frobenius norm (instead of the popularly used nuclear norm) is employed to enforce the graph to be more block-diagonal with lower complexity. Extensive experiments have been conducted on synthetic, real, and noisy data to show that the proposed method outperforms currently available alternatives by a margin of 1.0% 10.1%.
****
Towards High Fidelity Face Relighting With Realistic Shadows
Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14719-14728
Existing face relighting methods often struggle with two problems: maintaining the local facial details of the subject and accurately removing and synthesizing shadows in the relit image, especially hard shadows. We propose a novel deep face relighting method that addresses both problems. Our method learns to predict the ratio (quotient) image between a source image and the target image with the desired lighting, allowing us to relight the image while maintaining the local facial details. During training, our model also learns to accurately modify shadows by using estimated shadow masks to emphasize on the high-contrast shadow borders. Furthermore, we introduce a method to use the shadow mask to estimate the ambient light intensity in an image, and are thus able to leverage multiple datase

ts during training with different global lighting intensities. With quantitative and qualitative evaluations on the Multi-PIE and FFHQ datasets, we demonstrate that our proposed method faithfully maintains the local facial details of the subject and can accurately handle hard shadows while achieving state-of-the-art face relighting performance.
****
Multi-View Multi-Person 3D Pose Estimation With Plane Sweep Stereo
Jiahao Lin, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11886-11895
Existing approaches for multi-view multi-person 3D pose estimation explicitly establish cross-view correspondences to group 2D pose detections from multiple camera views and solve for the 3D pose estimation for each person. Establishing cross-view correspondences is challenging in multi-person scenes, and incorrect correspondences will lead to sub-optimal performance for the multi-stage pipeline. In this work, we present our multi-view 3D pose estimation approach based on plane sweep stereo to jointly address the cross-view fusion and 3D pose reconstruction in a single shot. Specifically, we propose to perform depth regression for each joint of each 2D pose in a target camera view. Cross-view consistency constraints are implicitly enforced by multiple reference camera views via the plane sweep algorithm to facilitate accurate depth regression. We adopt a coarse-to-fine scheme to first regress the person-level depth followed by a per-person joint-level relative depth estimation. 3D poses are obtained from a simple back-projection given the estimated depths. We evaluate our approach on benchmark datasets where it outperforms previous state-of-the-arts while being remarkably efficient.
****
Fusing the Old with the New: Learning Relative Camera Pose with Geometry-Guided Uncertainty
Bingbing Zhuang, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 32-42
Learning methods for relative camera pose estimation have been developed largely in isolation from classical geometric approaches. The question of how to integrate predictions from deep neural networks (DNNs) and solutions from geometric solvers, such as the 5-point algorithm, has as yet remained under-explored. In this paper, we present a novel framework that involves probabilistic fusion between the two families of predictions during network training, with a view to leveraging their complementary benefits in a learnable way. The fusion is achieved by learning the DNN uncertainty under explicit guidance by the geometric uncertainty, thereby learning to take into account the geometric solution in relation to the DNN prediction. Our network features a self-attention graph neural network, which drives the learning by enforcing strong interactions between different correspondences and potentially modeling complex relationships between points. We propose motion parmeterizations suitable for learning and show that our method achieves state-of-the-art performance on the challenging DeMoN and ScanNet datasets. While we focus on relative pose, we envision that our pipeline is broadly applicable for fusing classical geometry and deep learning.
****
CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning
Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, Fan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10857-10866
Semi-supervised learning on class-imbalanced data, although a realistic problem, has been under studied. While existing semi-supervised learning (SSL) methods are known to perform poorly on minority classes, we find that they still generate high precision pseudo-labels on minority classes. By exploiting this property, in this work, we propose Class-Rebalancing Self-Training (CReST), a simple yet effective framework to improve existing SSL methods on class-imbalanced data. CReST iteratively retrains a baseline SSL model with a labeled set expanded by adding pseudo-labeled samples from an unlabeled set, where pseudo-labeled samples fr

om minority classes are selected more frequently according to an estimated class distribution. We also propose a progressive distribution alignment to adaptively adjust the rebalancing strength dubbed CReST+. We show that CReST and CReST+ improve state-of-the-art SSL algorithms on various class-imbalanced datasets and consistently outperform other popular rebalancing methods. Code has been made available at https://github.com/google-research/crest.
****
Towards Diverse Paragraph Captioning for Untrimmed Videos
Yuqing Song, Shizhe Chen, Qin Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11245-11254
Video paragraph captioning aims to describe multiple events in untrimmed videos with descriptive paragraphs. Existing approaches mainly solve the problem in two steps: event detection and then event captioning. Such two-step manner makes the quality of generated paragraphs highly dependent on the accuracy of event proposal detection which is already a challenging task. In this paper, we propose a paragraph captioning model which eschews the problematic event detection stage and directly generates paragraphs for untrimmed videos. To describe coherent and diverse events, we propose to enhance the conventional temporal attention with dynamic video memories, which progressively exposes new video features and suppresses over-accessed video contents to control visual focuses of the model. In addition, a diversity-driven training strategy is proposed to improve diversity of paragraph on the language perspective. Considering that untrimmed videos generally contain massive but redundant frames, we further augment the video encoder with keyframe awareness to improve efficiency. Experimental results on the ActivityNet and Charades datasets show that our proposed model significantly outperforms the state-of-the-art performance on both accuracy and diversity metrics without using any event boundary annotations. Code will be released at https://github.com/syuqings/video-paragraph.
****
FlowStep3D: Model Unrolling for Self-Supervised Scene Flow Estimation
Yair Kittenplon, Yonina C. Eldar, Dan Raviv; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4114-4123
Estimating the 3D motion of points in a scene, known as scene flow, is a core problem in computer vision. Traditional learning-based methods designed to learn end-to-end 3D flow often suffer from poor generalization. Here we present a recurrent architecture that learns a single step of an unrolled iterative alignment procedure for refining scene flow predictions. Inspired by classical algorithms, we demonstrate iterative convergence toward the solution using strong regularization. The proposed method can handle sizeable temporal deformations and suggests a slimmer architecture than competitive all-to-all correlation approaches. Trained on FlyingThings3D synthetic data only, our network successfully generalizes to real scans, outperforming all existing methods by a large margin on the KITTI self-supervised benchmark.
****
Adversarial Robustness Across Representation Spaces
Pranjal Awasthi, George Yu, Chun-Sung Ferng, Andrew Tomkins, Da-Cheng Juan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7608-7616
Adversarial robustness corresponds to the susceptibility of deep neural networks to imperceptible perturbations made at test time. In the context of image tasks, many algorithms have been proposed to make neural networks robust to adversarial perturbations made to the input pixels. These perturbations are typically measured in an l_p norm. However, robustness often holds only for the specific attack used for training. In this work we extend the above setting to consider the problem of training of deep neural networks that can be made simultaneously robust to perturbations applied in multiple natural representations spaces. For the case of image data, examples include the standard pixel representation as well as the representation in the discrete cosine transform (DCT) basis. We design a theoretically sound algorithm with formal guarantees for the above problem. Furthermore, our guarantees also hold when the goal is to require robustness with resp

ect to multiple l_p norm based attacks. We then derive an efficient practical im
plementation and demonstrate the effectiveness of our approach on standard datas
ets for image classification.
****
MagDR: Mask-Guided Detection and Reconstruction for Defending Deepfakes
Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, Bo Zhang; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 90
14-9023
Deepfakes raised serious concerns on the authenticity of visual contents. Prior
works revealed the possibility to disrupt deepfakes by adding adversarial pertur
bations to the source data, but we argue that the threat has not been eliminated
 yet. This paper presents MagDR, a mask-guided detection and reconstruction pipe
line for defending deepfakes from adversarial attacks. MagDR starts with a detec
tion module that defines a few criteria to judge the abnormality of the output o
f deepfakes, and then uses it to guide an learnable reconstruction procedure. Ad
aptive masks are extracted to capture the change in local facial regions. In exp
eriments, MagDR defends three main tasks of deepfakes, and the learned reconstru
ction pipeline transfers across input data, showing promising performance in def
ending both black-box and white-box attacks.
****
Neural Deformation Graphs for Globally-Consistent Non-Rigid Reconstruction
Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, Matthia
s Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2021, pp. 1450-1459
We introduce Neural Deformation Graphs for globally-consistent deformation track
ing and 3D reconstruction of non-rigid objects. Specifically, we implicitly mode
l a deformation graph via a deep neural network. This neural deformation graph d
oes not rely on any object-specific structure and, thus, can be applied to gener
al non-rigid deformation tracking. Our method globally optimizes this neural gra
ph on a given sequence of depth camera observations of a non-rigidly moving obje
ct. Based on explicit viewpoint consistency as well as inter-frame graph and sur
face consistency constraints, the underlying network is trained in a self-superv
ised fashion. We additionally optimize for the geometry of the object with an im
plicit deformable multi-MLP shape representation. Our approach does not assume s
equential input data, thus enabling robust tracking of fast motions or even temp
orally disconnected recordings. Our experiments demonstrate that our Neural Defo
rmation Graphs outperform state-of-the-art non-rigid reconstruction approaches b
oth qualitatively and quantitatively, with 64% improved reconstruction and 54% i
mproved deformation tracking performance. Code is publicly available.
****
Fostering Generalization in Single-View 3D Reconstruction by Learning a Hierarch
y of Local and Global Shape Priors
Jan Bechtold, Maxim Tatarchenko, Volker Fischer, Thomas Brox; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp
. 15880-15889
Single-view 3D object reconstruction has seen much progress, yet methods still s
truggle generalizing to novel shapes unseen during training. Common approaches p
redominantly rely on learned global shape priors and, hence, disregard detailed
local observations. In this work, we address this issue by learning a hierarchy
of priors at different levels of locality from ground truth input depth maps. We
 argue that exploiting local priors allows our method to efficiently use input o
bservations, thus improving generalization in visible areas of novel shapes. At
the same time, the combination of local and global priors enables meaningful hal
lucination of unobserved parts resulting in consistent 3D shapes. We show that t
he hierarchical approach generalizes much better than the global approach. It ge
neralizes not only between different instances of a class but also across classe
s and to unseen arrangements of objects.
****
Progressive Semantic-Aware Style Transformation for Blind Face Restoration
Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, Kwan-Yee K. Won

g; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11896-11905

Face restoration is important in face image processing, and has been widely studied in recent years. However, previous works often fail to generate plausible high quality (HQ) results for real-world low quality (LQ) face images. In this paper, we propose a new progressive semantic-aware style transformation framework, named PSFR-GAN, for face restoration. Specifically, instead of using an encoder-decoder framework as previous methods, we formulate the restoration of LQ face images as a multi-scale progressive restoration procedure through semantic-aware style transformation. Given a pair of LQ face image and its corresponding parsing map, we first generate a multi-scale pyramid of the inputs, and then progressively modulate different scale features from coarse-to-fine in a semantic-aware style transfer way. Compared with previous networks, the proposed PSFR-GAN makes full use of the semantic (parsing maps) and pixel (LQ images) space information from different scales of input pairs. In addition, we further introduce a semantic aware style loss which calculates the feature style loss for each semantic region individually to improve the details of face textures. Finally, we pretrain a face parsing network which can generate decent parsing maps from real-world LQ face images. Experiment results show that our model trained with synthetic data can produce more realistic high-resolution results for synthetic LQ inputs than state-of-the-art methods and generalize better to natural LQ face images.

****

Seeking the Shape of Sound: An Adaptive Framework for Learning Voice-Face Association

Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, Qingming Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16347-16356

Nowadays, we have witnessed the early progress on learning the association between voice and face automatically, which brings a new wave of studies to the computer vision community. However, most of the prior arts along this line (a) merely adopt local information to perform modality alignment and (b) ignore the diversity of learning difficulty across different subjects. In this paper, we propose a novel framework to jointly address the above-mentioned issues. Targeting at (a), we propose a two-level modality alignment loss where both global and local information are considered. Compared with the existing methods, we introduce a global loss into the modality alignment process. The global component of the loss is driven by the accuracy of the identity classification. Theoretically, we show that minimizing the loss could maximize the distance between embeddings across different identities while minimizing the distance between embeddings belonging to the same identity, in a global sense (instead of a mini-batch). Targeting at (b), we propose a dynamic reweighting scheme to better explore the hard but valuable identities while filtering out the unlearnable and noisy identities. Experiments show that the proposed method outperforms the previous methods in multiple settings, including voice-face matching, verification and retrieval.

****

Invertible Image Signal Processing

Yazhou Xing, Zian Qian, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6287-6296

Unprocessed RAW data is a highly valuable image format for image editing and computer vision. However, since the file size of RAW data is huge, most users can only get access to processed and compressed sRGB images. To bridge this gap, we design an Invertible Image Signal Processing (InvISP) pipeline, which not only enables rendering visually appealing sRGB images but also allows recovering nearly perfect RAW data. Due to our framework's inherent reversibility, we can reconstruct realistic RAW data instead of synthesizing RAW data from sRGB images without any memory overhead. We also integrate a differentiable JPEG compression simulator that empowers our framework to reconstruct RAW data from JPEG images. Extensive quantitative and qualitative experiments on two DSLR demonstrate that our method obtains much higher quality in both rendered sRGB images and reconstructed RAW data than alternative methods.

****
Lighting, Reflectance and Geometry Estimation From 360deg Panoramic Stereo
Junxuan Li, Hongdong Li, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10591-10600
We propose a method for estimating high-definition spatially-varying lighting, reflectance, and geometry of a scene from 360deg stereo images. Our model takes advantage of the 360deg input to observe the entire scene with geometric detail, then jointly estimates the scene's properties with physical constraints. We first reconstruct a near-field environment light for predicting the lighting at any 3D location within the scene. Then we present a deep learning model that leverages the stereo information to infer the reflectance and surface normal. Lastly, we incorporate the physical constraints between lighting and geometry to refine the reflectance of the scene. Both quantitative and qualitative experiments show that our method, benefiting from the 360deg observation of the scene, outperforms prior state-of-the-art methods and enables more augmented reality applications such as mirror-objects insertion.
****
Building Reliable Explanations of Unreliable Neural Networks: Locally Smoothing Perspective of Model Interpretation
Dohun Lim, Hyeonseok Lee, Sungchan Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6468-6477
We present a novel method for reliably explaining the predictions of neural networks. We consider an explanation reliable if it identifies input features relevant to the model output by considering the input and the neighboring data points. Our method is built on top of the assumption of smooth landscape in a loss function of the model prediction: locally consistent loss and gradient profile. A theoretical analysis established in this study suggests that those locally smooth model explanations are learned using a batch of noisy copies of the input with the L1 regularization for a saliency map. Extensive experiments support the analysis results, revealing that the proposed saliency maps retrieve the original classes of adversarial examples crafted against both naturally and adversarially trained models, significantly outperforming previous methods. We further demonstrated that such good performance results from the learning capability of this method to identify input features that are truly relevant to the model output of the input and the neighboring data points, fulfilling the requirements of a reliable explanation.
****
NeX: Real-Time View Synthesis With Neural Basis Expansion
Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8534-8543
We present NeX, a new approach to novel view synthesis based on enhancements of multiplane image (MPI) that can reproduce next-level view-dependent effects--in real time. Unlike traditional MPI that uses a set of simple RGBa planes, our technique models view-dependent effects by instead parameterizing each pixel as a linear combination of basis functions learned from a neural network. Moreover, we propose a hybrid implicit-explicit modeling strategy that improves upon fine detail and produces state-of-the-art results. Our method is evaluated on benchmark forward-facing datasets as well as our newly-introduced dataset designed to test the limit of view-dependent modeling with significantly more challenging effects such as the rainbow reflections on a CD. Our method achieves the best overall scores across all major metrics on these datasets with more than 1000x faster rendering time than the state of the art. For real-time demos, visit https://nex-mpi.github.io/
****
DAT: Training Deep Networks Robust To Label-Noise by Matching the Feature Distributions
Yuntao Qu, Shasha Mo, Jianwei Niu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6821-6829
In real application scenarios, the performance of deep networks may be degraded

when the dataset contains noisy labels. Existing methods for learning with noisy labels are limited by two aspects. Firstly, methods based on the noise probability modeling can only be applied to class-level noisy labels. Secondly, others based on the memorization effect outperform in synthetic noise but get weak promotion in real-world noisy datasets. To solve these problems, this paper proposes a novel label-noise robust method named Discrepant Adversarial Training (DAT). The DAT method has ability of enforcing prominent feature extraction by matching feature distribution between clean and noisy data. Therefore, under the noise-free feature representation, the deep network can simply output the correct result. To better capture the divergence between the noisy and clean distribution, a new metric is designed to change the distribution divergence into computable. By minimizing the proposed metric with a min-max training of discrepancy on classifiers and generators, DAT can match noisy data to clean data in the feature space. To the best of our knowledge, DAT is the first to address the noisy label problem from the perspective of the feature distribution. Experiments on synthetic and real-world noisy datasets demonstrate that DAT can consistently outperform other state-of-the-art methods. Codes are available at https://github.com/Tyqnn0323/DAT.
****
Repetitive Activity Counting by Sight and Sound
Yunhua Zhang, Ling Shao, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14070-14079
This paper strives for repetitive activity counting in videos. Different from existing works, which all analyze the visual video content only, we incorporate for the first time the corresponding sound into the repetition counting process. This benefits accuracy in challenging vision conditions such as occlusion, dramatic camera view changes, low resolution, etc. We propose a model that starts with analyzing the sight and sound streams separately. Then an audiovisual temporal stride decision module and a reliability estimation module are introduced to exploit cross-modal temporal interaction. For learning and evaluation, an existing dataset is repurposed and reorganized to allow for repetition counting with sight and sound. We also introduce a variant of this dataset for repetition counting under challenging vision conditions. Experiments demonstrate the benefit of sound, as well as the other introduced modules, for repetition counting. Our sight-only model already outperforms the state-of-the-art by itself, when we add sound, results improve notably, especially under harsh vision conditions.
****
PointGuard: Provably Robust 3D Point Cloud Classification
Hongbin Liu, Jinyuan Jia, Neil Zhenqiang Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6186-6195
3D point cloud classification has many safety-critical applications such as autonomous driving and robotic grasping. However, several studies showed that it is vulnerable to adversarial attacks. In particular, an attacker can make a classifier predict an incorrect label for a 3D point cloud via carefully modifying, adding, and/or deleting a small number of its points. Randomized smoothing is state-of-the-art technique to build certifiably robust 2D image classifiers. However, when applied to 3D point cloud classification, randomized smoothing can only certify robustness against adversarially modified points. In this work, we propose PointGuard, the first defense that has provable robustness guarantees against adversarially modified, added, and/or deleted points. Specifically, given a 3D point cloud and an arbitrary point cloud classifier, our PointGuard first creates multiple subsampled point clouds, each of which contains a random subset of the points in the original point cloud; then our PointGuard predicts the label of the original point cloud as the majority vote among the labels of the subsampled point clouds predicted by the point cloud classifier. Our first major theoretical contribution is that we show PointGuard provably predicts the same label for a 3D point cloud when the number of adversarially modified, added, and/or deleted points is bounded. Our second major theoretical contribution is that we prove the tightness of our derived bound when no assumptions on the point cloud classifier are made. Moreover, we design an efficient algorithm to compute our certified

robustness guarantees. We also empirically evaluate PointGuard on ModelNet40 and ScanNet benchmark datasets.
****

Unsupervised Multi-Source Domain Adaptation for Person Re-Identification
Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12914-12923
Unsupervised domain adaptation (UDA) methods for person re-identification (re-ID) aim at transferring re-ID knowledge from labeled source data to unlabeled target data. Among these methods, the pseudo-label-based branch has achieved great success, whereas most of them only use limited data from a single-source domain for model pre-training, making the rich labeled data insufficiently exploited. To make full use of the valuable labeled data, we introduce the multi-source concept into UDA person re-ID field, where multiple source datasets are used during training. However, because of domain gaps, simply combining different datasets only brings limited improvement. In this paper, we try to address this problem from two perspectives, i.e. domain-specific view and domain-fusion view. Two constructive modules are proposed, and they are compatible with each other. First, a rectification domain-specific batch normalization (RDSBN) module is explored to simultaneously reduce domain-specific characteristics and increase the distinctiveness of person features. Second, a graph convolutional network (GCN) based multi-domain information fusion (MDIF) module is developed, which minimizes domain distances by fusing features of different domains. The proposed method outperforms state-of-the-art UDA person re-ID methods by a large margin, and even achieves comparable performance to the supervised approaches without any post-processing techniques.
****

BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation
Jungbeom Lee, Jihun Yi, Chaehun Shin, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2643-2652
Weakly supervised segmentation methods using bounding box annotations focus on obtaining a pixel-level mask from each box containing an object. Existing methods typically depend on a class-agnostic mask generator, which operates on the low-level information intrinsic to an image. In this work, we utilize higher-level information from the behavior of a trained object detector, by seeking the smallest areas of the image from which the object detector produces almost the same result as it does from the whole image. These areas constitute a bounding-box attribution map (BBAM), which identifies the target object in its bounding box and thus serves as pseudo ground-truth for weakly supervised semantic and instance segmentation. This approach significantly outperforms recent comparable techniques on both the PASCAL VOC and MS COCO benchmarks in weakly supervised semantic and instance segmentation. In addition, we provide a detailed analysis of our method, offering deeper insight into the behavior of the BBAM.
****

Boosting Video Representation Learning With Multi-Faceted Integration
Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xiao-Ping Zhang, Dong Wu, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14030-14039
Video content is multifaceted, consisting of objects, scenes, interactions or actions. The existing datasets mostly label only one of the facets for model training, resulting in the video representation that biases to only one facet depending on the training dataset. There is no study yet on how to learn a video representation from multifaceted labels, and whether multifaceted information is helpful for video representation learning. In this paper, we propose a new learning framework, MUlti-Faceted Integration (MUFI), to aggregate facets from different datasets for learning a representation that could reflect the full spectrum of video content. Technically, MUFI formulates the problem as visual-semantic embedding learning, which explicitly maps video representation into a rich semantic emb

edding space, and jointly optimizes video representation from two perspectives. One is to capitalize on the intra-facet supervision between each video and its own label descriptions, and the second predicts the "semantic representation" of each video from the facets of other datasets as the inter-facet supervision. Extensive experiments demonstrate that learning 3D CNN via our MUFI framework on a union of four large-scale video datasets plus two image datasets leads to superior capability of video representation. The pre-learnt 3D CNN with MUFI also shows clear improvements over other approaches on several downstream video applications. More remarkably, MUFI achieves 98.1%/80.9% on UCF101/HMDB51 for action recognition and 101.5% in terms of CIDEr-D score on MSVD for video captioning.
****
Beyond Bounding-Box: Convex-Hull Feature Adaptation for Oriented and Densely Packed Object Detection
Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8792-8801
Detecting oriented and densely packed objects remains challenging for spatial feature aliasing caused by the intersection of reception fields between objects. In this paper, we propose a convex-hull feature adaptation (CFA) approach for configuring convolutional features in accordance with oriented and densely packed object layouts. CFA is rooted in convex-hull feature representation, which defines a set of dynamically predicted feature points guided by the convex intersection over union (CIoU) to bound the extent of objects. CFA pursues optimal feature assignment by constructing convex-hull sets and dynamically splitting positive or negative convex-hulls. By simultaneously considering overlapping convex-hulls and objects and penalizing convex-hulls shared by multiple objects, CFA alleviates spatial feature aliasing towards optimal feature adaptation. Experiments on DOTA and SKU110K-R datasets show that CFA significantly outperforms the baseline approach, achieving new state-of-the-art detection performance.
****
3D Graph Anatomy Geometry-Integrated Network for Pancreatic Mass Segmentation, Diagnosis, and Quantitative Patient Management
Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, Ling Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13743-13752
The pancreatic disease taxonomy includes ten types of masses (tumors or cysts) [20, 8]. Previous work focuses on developing segmentation or classification methods only for certain mass types. Differential diagnosis of all mass types is clinically highly desirable [20] but has not been investigated using an automated image understanding approach. We exploit the feasibility to distinguish pancreatic ductal adenocarcinoma (PDAC) from the nine other nonPDAC masses using multi-phase CT imaging. Both image appearance and the 3D organ-mass geometry relationship are critical. We propose a holistic segmentation-mesh-classification network (SMCN) to provide patient-level diagnosis, by fully utilizing the geometry and location information, which is accomplished by combining the anatomical structure and the semantic detection-by-segmentation network. SMCN learns the pancreas and mass segmentation task and builds an anatomical correspondence-aware organ mesh model by progressively deforming a pancreas prototype on the raw segmentation mask (i.e., mask-to-mesh). A new graph-based residual convolutional network (Graph-ResNet), whose nodes fuse the information of the mesh model and feature vectors extracted from the segmentation network, is developed to produce the patient-level differential classification results. Extensive experiments on 661 patients' CT scans (five phases per patient) show that SMCN can improve the mass segmentation and detection accuracy compared to the strong baseline method nnUNet (e.g., for nonPDAC, Dice: 0.611 vs. 0.478; detection rate: 89% vs. 70%), achieve similar sensitivity and specificity in differentiating PDAC and nonPDAC as expert radiologists (i.e., 94% and 90%), and obtain results comparable to a multimodality test [20] that combines clinical, imaging, and molecular testing for clinical management of patients.
****

Protecting Intellectual Property of Generative Adversarial Networks From Ambigui
ty Attacks
Ding Sheng Ong, Chee Seng Chan, Kam Woh Ng, Lixin Fan, Qiang Yang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 3630-3639
Ever since Machine Learning as a Service emerges as a viable business that utili
zes deep learning models to generate lucrative revenue, Intellectual Property Ri
ght (IPR) has become a major concern because these deep learning models can easi
ly be replicated, shared, and re-distributed by any unauthorized third parties.
To the best of our knowledge, one of the prominent deep learning models - Genera
tive Adversarial Networks (GANs) which has been widely used to create photoreali
stic image are totally unprotected despite the existence of pioneering IPR prote
ction methodology for Convolutional Neural Networks (CNNs). This paper therefore
presents a complete protection framework in both black-box and white-box settin
gs to enforce IPR protection on GANs. Empirically, we show that the proposed met
hod does not compromise the original GANs performance (i.e. image generation, im
age super-resolution, style transfer), and at the same time, it is able to withs
tand both removal and ambiguity attacks against embedded watermarks.
****
End-to-End High Dynamic Range Camera Pipeline Optimization
Nicolas Robidoux, Luis E. Garcia Capel, Dong-eun Seo, Avinash Sharma, Federico A
riza, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2021, pp. 6297-6307
With a 280 dB dynamic range, the real world is a High Dynamic Range (HDR) world.
Today's sensors cannot record this dynamic range in a single shot. Instead, HDR
cameras acquire multiple measurements with different exposures, gains and photo
diodes, from which an Image Signal Processor (ISP) reconstructs an HDR image. HD
R image recovery for dynamic scenes is an open challenge because of motion and b
ecause stitched captures have different noise characteristics, resulting in arte
facts that the ISP has to resolve---in real time and at triple-digit megapixel r
esolutions. Traditionally, hardware ISP settings used by downstream vision modul
es have been chosen by domain experts. Such frozen camera designs are then used
for training data acquisition and supervised learning of downstream vision modul
es. We depart from this paradigm and formulate HDR ISP hyperparameter search as
an end-to-end optimization problem. We propose a mixed 0th and 1st-order block c
oordinate descent optimizer to jointly learn ISP and detector network weights us
ing RAW image data augmented with emulated SNR transition region artefacts. We a
ssess the proposed method for human vision and image understanding. For automoti
ve object detection, the method improves mAP and mAR by 33% compared to expert-t
uning and by 22% compared to recent state-of-the-art. The method is validated in
an HDR laboratory rig and in the field, outperforming conventional handcrafted
HDR imaging and vision pipelines in all experiments.
****
Parser-Free Virtual Try-On via Distilling Appearance Flows
Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, Ping Luo; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2021, pp. 8485-8493
Image virtual try-on aims to fit a garment image (target clothes) to a person im
age. Prior methods are heavily based on human parsing. However, slightly-wrong s
egmentation results would lead to unrealistic try-on images with large artifacts
. Inaccurate parsing misleads parser-based methods to produce visually unrealist
ic results where artifacts usually occur. A recent pioneering work employed know
ledge distillation to reduce the dependency of human parsing, where the try-on i
mages produced by a parser-based method are used as supervisions to train a "stu
dent" network without relying on segmentation, making the student mimic the try-
on ability of the parser-based model. However, the image quality of the student
is bounded by the parser-based model. To address this problem, we propose a nove
l approach, "teacher-tutor-student" knowledge distillation, which is able to pro
duce highly photo-realistic images without human parsing, possessing several app
ealing advantages compared to prior arts. (1) Unlike existing work, our approach

treats the fake images produced by the parser-based method as "tutor knowledge" , where the artifacts can be corrected by real "teacher knowledge", which is extracted from the real person images in a self-supervised way. (2) Other than using real images as supervisions, we formulate knowledge distillation in the try-on problem as distilling the appearance flows between the person image and the garment image, enabling us to find accurate dense correspondences between them to produce high-quality results. (3) Extensive evaluations show large superiority of our method (see Fig. 1).
****

GIRAFFE: Representing Scenes As Compositional Generative Neural Feature Fields
Michael Niemeyer, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11453-11464
Deep generative models allow for photorealistic image synthesis at high resolutions. But for many applications, this is not enough: content creation also needs to be controllable. While several recent works investigate how to disentangle underlying factors of variation in the data, most of them operate in 2D and hence ignore that our world is three-dimensional. Further, only few works consider the compositional nature of scenes. Our key hypothesis is that incorporating a compositional 3D scene representation into the generative model leads to more controllable image synthesis. Representing scenes as compositional generative neural feature fields allows us to disentangle one or multiple objects from the background as well as individual objects' shapes and appearances while learning from unstructured and unposed image collections without any additional supervision. Combining this scene representation with a neural rendering pipeline yields a fast and realistic image synthesis model. As evidenced by our experiments, our model is able to disentangle individual objects and allows for translating and rotating them in the scene as well as changing the camera pose.
****

Single-Stage Instance Shadow Detection With Bidirectional Relation Learning
Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, Pheng-Ann Heng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1-11
Instance shadow detection aims to find shadow instances paired with the objects that cast the shadows. The previous work adopts a two-stage framework to first predict shadow instances, object instances, and shadow-object associations from the region proposals, then leverage a post-processing to match the predictions to form the final shadow-object pairs. In this paper, we present a new single-stage fully-convolutional network architecture with a bidirectional relation learning module to directly learn the relations of shadow and object instances in an end-to-end manner. Compared with the prior work, our method actively explores the internal relationship between shadows and objects to learn a better pairing between them, thus improving the overall performance for instance shadow detection. We evaluate our method on the benchmark dataset for instance shadow detection, both quantitatively and visually. The experimental results demonstrate that our method clearly outperforms the state-of-the-art method.
****

High-Speed Image Reconstruction Through Short-Term Plasticity for Spiking Cameras
Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6358-6367
Fovea, located in the centre of the retina, is specialized for high-acuity vision. Mimicking the sampling mechanism of the fovea, a retina-inspired camera, named spiking camera, is developed to record the external information with a sampling rate of 40,000 Hz, and outputs asynchronous binary spike streams. Although the temporal resolution of visual information is improved, how to reconstruct the scenes is still a challenging problem. In this paper, we present a novel high-speed image reconstruction model through the short-term plasticity (STP) mechanism of the brain. We derive the relationship between postsynaptic potential regulated by STP and the firing frequency of each pixel. By setting up the STP model at each pixel of the spiking camera, we can infer the scene radiance with the tempo

ral regularity of the spike stream. Moreover, we show that STP can be used to distinguish the static and motion areas and further enhance the reconstruction results. The experimental results show that our methods achieve state-of-the-art performance in both image quality and computing time.
****
Self-Supervised 3D Mesh Reconstruction From Single Images
Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6002-6011
Recent single-view 3D reconstruction methods reconstruct object's shape and texture from a single image with only 2D image-level annotation. However, without explicit 3D attribute-level supervision, it is still difficult to achieve satisfying reconstruction accuracy. In this paper, we propose a Self-supervised Mesh Reconstruction (SMR) approach to enhance 3D mesh attribute learning process. Our approach is motivated by observations that (1) 3D attributes from interpolation and prediction should be consistent, and (2) feature representation of landmarks from all images should be consistent. By only requiring silhouette mask annotation, our SMR can be trained in an end-to-end manner and generalizes to reconstruct natural objects of birds, cows, motorbikes, etc. Experiments demonstrate that our approach improves both 2D supervised and unsupervised 3D mesh reconstruction on multiple datasets. We also show that our model can be adapted to other image synthesis tasks, e.g., novel view generation, shape transfer, and texture transfer, with promising results. Our code is publicly available at https://github.com/Jia-Research-Lab.
****
Dual-GAN: Joint BVP and Noise Modeling for Remote Physiological Measurement
Hao Lu, Hu Han, S. Kevin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12404-12413
Remote photoplethysmography (rPPG) based physiological measurement has great application values in health monitoring, emotion analysis, etc. Existing methods mainly focus on how to enhance or extract the very weak blood volume pulse (BVP) signals from face videos, but seldom explicitly model the noises that dominate face video content. Thus, they may suffer from poor generalization ability in unseen scenarios. This paper proposes a novel adversarial learning approach for rPPG based physiological measurement by using Dual Generative Adversarial Networks (Dual-GAN) to model the BVP estimation and noise distribution jointly. The BVP-GAN aims to learn a noise-resistant mapping from input to ground-truth BVP, and the Noise-GAN aims to learn the noise distribution. The dual GANs can promote each other's capability, leading to improved feature disentanglement between BVP and noises. Besides, a plug-and-play block named ROI alignment and fusion (ROI-AF) block is proposed to alleviate the inconsistencies between different ROIs and exploit informative features from a wider receptive field in terms of ROIs. In comparison to state-of-the-art methods, our method achieves better performance in heart rate, heart rate variability, and respiration frequency estimation from face videos.
****
Audio-Visual Instance Discrimination with Cross-Modal Agreement
Pedro Morgado, Nuno Vasconcelos, Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12475-12486
We present a self-supervised learning approach to learn audio-visual representations from video and audio. Our method uses contrastive learning for cross-modal discrimination of video from audio and vice-versa. We show that optimizing for cross-modal discrimination, rather than within-modal discrimination, is important to learn good representations from video and audio. With this simple but powerful insight, our method achieves highly competitive performance when finetuned on action recognition tasks. Furthermore, while recent work in contrastive learning defines positive and negative samples as individual instances, we generalize this definition by exploring cross-modal agreement. We group together multiple instances as positives by measuring their similarity in both the video and audio feature spaces. Cross-modal agreement creates better positive and negative sets,

which allows us to calibrate visual similarities by seeking within-modal discrimination of positive instances, and achieve significant gains on downstream tasks.

****

Combined Depth Space Based Architecture Search for Person Re-Identification

Hanjun Li, Gaojie Wu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6729-6738

Most works on person re-identification (ReID) take advantage of large backbone networks such as ResNet, which are designed for image classification instead of ReID, for feature extraction. However, these backbones may not be computationally efficient or the most suitable architectures for ReID. In this work, we aim to design a lightweight and suitable network for ReID. To this end, we propose a novel search space called Combined Depth Space (CDS), based on which we search for an efficient network architecture, which we call CDNet, via a differentiable architecture search algorithm. Through the use of the combined basic building blocks in CDS, CDNet tends to focus on combined pattern information that is typically found in images of pedestrians. We then propose a low-cost search strategy named the Top-k Sample Search strategy to make full use of the search space and avoid trapping in local optimal result. Furthermore, an effective Fine-grained Balance Neck (FBLNeck), which is removable at the inference time, is presented to balance the effects of triplet loss and softmax loss during the training process. Extensive experiments show that our CDNet (1.8 M parameters) has comparable performance with state-of-the-art lightweight networks.

****

Rethinking BiSeNet for Real-Time Semantic Segmentation

Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, Xiaolin Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9716-9725

BiSeNet has been proved to be a popular two-stream network for real-time segmentation. However, its principle of adding an extra path to encode spatial information is time-consuming, and the backbones borrowed from pretrained tasks, e.g., image classification, may be inefficient for image segmentation due to the deficiency of task-specific design. To handle these problems, we propose a novel and efficient structure named Short-Term Dense Concatenate network (STDC network) by removing structure redundancy. Specifically, we gradually reduce the dimension of feature maps and use the aggregation of them for image representation, which forms the basic module of STDC network. In the decoder, we propose a Detail Aggregation module by integrating the learning of spatial information into low-level layers in single-stream manner. Finally, the low-level features and deep features are fused to predict the final segmentation results. Extensive experiments on Cityscapes and CamVid dataset demonstrate the effectiveness of our method by achieving promising trade-off between segmentation accuracy and inference speed. On Cityscapes, we achieve 71.9% mIoU on the test set with a speed of 250.4 FPS on NVIDIA GTX 1080Ti, which is 45.2% faster than the latest methods, and achieve 76.8% mIoU with 97.0 FPS while inferring on higher resolution images.

****

The Spatially-Correlative Loss for Various Image Translation Tasks

Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16407-16417

We propose a novel spatially-correlative loss that is simple, efficient, and yet effective for preserving scene structure consistency while supporting large appearance changes during unpaired image-to-image (I2I) translation. Previous methods attempt this by using pixel-level cycle-consistency or feature-level matching losses, but the domain-specific nature of these losses hinder translation across large domain gaps. To address this, we exploit the spatial patterns of self-similarity as a means of defining scene structure. Our spatially-correlative loss is geared towards only capturing spatial relationships within an image rather than domain appearance. We also introduce a new self-supervised learning method to explicitly learn spatially-correlative maps for each specific translation task. We show distinct improvement over baseline models in all three modes of unpaire

d I2I translation: single-modal, multi-modal, and even single-image translation. This new loss can easily be integrated into existing network architectures and thus allows wide applicability.
****
Learning To Restore Hazy Video: A New Real-World Dataset and a New Method
Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Fei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9239-9248
Most of the existing deep learning-based dehazing methods are trained and evaluated on the image dehazing datasets, where the dehazed images are generated by only exploiting the information from the corresponding hazy ones. On the other hand, the video dehazing algorithms, which can acquire more satisfying dehazing results by exploiting the temporal redundancy from neighborhood hazy frames, receive less attention due to the absence of the video dehazing datasets. Therefore, we propose the first REal-world VIdeo DEhazing (REVIDE) dataset which can be used for the supervised learning of the video dehazing algorithms. By utilizing a well-designed video acquisition system, we can capture paired real-world hazy and haze-free videos that are perfectly aligned by recording the same scene (with or without haze) twice. Considering the challenge of exploiting temporal redundancy among the hazy frames, we also develop a Confidence Guided and Improved Deformable Network (CG-IDN) for video dehazing. The experiments demonstrate that the hazy scenes in the REVIDE dataset are more realistic than the synthetic datasets and the proposed algorithm also performs favorably against state-of-the-art dehazing methods.
****
DyGLIP: A Dynamic Graph Model With Link Prediction for Accurate Multi-Camera Multiple Object Tracking
Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13784-13793
Multi-Camera Multiple Object Tracking (MC-MOT) is a significant computer vision problem due to its emerging applicability in several real-world applications. Despite a large number of existing works, solving the data association problem in any MC-MOT pipeline is arguably one of the most challenging tasks. Developing a robust MC-MOT system, however, is still highly challenging due to many practical issues such as inconsistent lighting conditions, varying object movement patterns, or the trajectory occlusions of the objects between the cameras. To address these problems, this work, therefore, proposes a new Dynamic Graph Model with Link Prediction (DyGLIP) approach to solve the data association task. Compared to existing methods, our new model offers several advantages, including better feature representations and the ability to recover from lost tracks during camera transitions. Moreover, our model works gracefully regardless of the overlapping ratios between the cameras. Experimental results show that we outperform existing MC-MOT algorithms by a large margin on several practical datasets. Notably, our model works favorably on online settings but can be extended to an incremental approach for large-scale datasets.
****
Towards Efficient Tensor Decomposition-Based DNN Model Compression With Optimization Framework
Miao Yin, Yang Sui, Siyu Liao, Bo Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10674-10683
Advanced tensor decomposition, such as Tensor train (TT) and Tensor ring (TR), has been widely studied for deep neural network (DNN) model compression, especially for recurrent neural networks (RNNs). However, compressing convolutional neural networks (CNNs) using TT/TR always suffers significant accuracy loss. In this paper, we propose a systematic framework for tensor decomposition-based model compression using Alternating Direction Method of Multipliers (ADMM). By formulating TT decomposition-based model compression to an optimization problem with constraints on tensor ranks, we leverage ADMM technique to systemically solve this optimization problem in an iterative way. During this procedure, the entire DNN

model is trained in the original structure instead of TT format, but gradually enjoys the desired low tensor rank characteristics. We then decompose this uncompressed model to TT format and fine-tune it to finally obtain a high-accuracy TT-format DNN model. Our framework is very general, and it works for both CNNs and RNNs, and can be easily modified to fit other tensor decomposition approaches. We evaluate our proposed framework on different DNN models for image classification and video recognition tasks. Experimental results show that our ADMM-based TT-format models demonstrate very high compression performance with high accuracy. Notably, on CIFAR-100, with 2.3X and 2.4X compression ratios, our models have 1.96% and 2.21% higher top-1 accuracy than the original ResNet-20 and ResNet-32, respectively. For compressing ResNet-18 on ImageNet, our model achieves 2.47X FLOPs reduction without accuracy loss.
****
User-Guided Line Art Flat Filling With Split Filling Mechanism
Lvmin Zhang, Chengze Li, Edgar Simo-Serra, Yi Ji, Tien-Tsin Wong, Chunping Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9889-9898
Flat filling is a critical step in digital artistic content creation with the objective of filling line arts with flat colors. We present a deep learning framework for user-guided line art flat filling that can compute the "influence areas" of the user color scribbles, i.e., the areas where the user scribbles should propagate and influence. This framework explicitly controls such scribble influence areas for artists to manipulate the colors of image details and avoid color leakage/contamination between scribbles, and simultaneously, leverages data-driven color generation to facilitate content creation. This framework is based on a Split Filling Mechanism (SFM), which first splits the user scribbles into individual groups and then independently processes the colors and influence areas of each group with a Convolutional Neural Network (CNN). Learned from more than a million illustrations, the framework can estimate the scribble influence areas in a content-aware manner, and can smartly generate visually pleasing colors to assist the daily works of artists. We show that our proposed framework is easy to use, allowing even amateurs to obtain professional-quality results on a wide variety of line arts.
****
Restore From Restored: Video Restoration With Pseudo Clean Video
Seunghwan Lee, Donghyeon Cho, Jiwon Kim, Tae Hyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3537-3546
In this study, we propose a self-supervised video denoising method called ""restore-from-restored."" This method fine-tunes a pre-trained network by using a pseudo clean video during the test phase. The pseudo clean video is obtained by applying a noisy video to the baseline network. By adopting a fully convolutional neural network (FCN) as the baseline, we can improve video denoising performance without accurate optical flow estimation and registration steps, in contrast to many conventional video restoration methods, due to the translation equivariant property of the FCN. Specifically, the proposed method can take advantage of plentiful similar patches existing across multiple consecutive frames (i.e., patch-recurrence); these patches can boost the performance of the baseline network by a large margin. We analyze the restoration performance of the fine-tuned video denoising networks with the proposed self-supervision-based learning algorithm, and demonstrate that the FCN can utilize recurring patches without requiring accurate registration among adjacent frames. In our experiments, we apply the proposed method to state-of-the-art denoisers and show that our fine-tuned networks achieve a considerable improvement in denoising performance.
****
Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion
Shi Qiu, Saeed Anwar, Nick Barnes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1757-1767
Given the prominence of current 3D sensors, a fine-grained analysis on the basic

point cloud data is worthy of further investigation. Particularly, real point cloud scenes can intuitively capture complex surroundings in the real world, but due to 3D data's raw nature, it is very challenging for machine perception. In this work, we concentrate on the essential visual task, semantic segmentation, for large-scale point cloud data collected in reality. On the one hand, to reduce the ambiguity in nearby points, we augment their local context by fully utilizing both geometric and semantic features in a bilateral structure. On the other hand, we comprehensively interpret the distinctness of the points from multiple resolutions and represent the feature map following an adaptive fusion method at point-level for accurate semantic segmentation. Further, we provide specific ablation studies and intuitive visualizations to validate our key modules. By comparing with state-of-the-art networks on three different benchmarks, we demonstrate the effectiveness of our network.
****
Interactive Self-Training With Mean Teachers for Semi-Supervised Object Detection
Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5941-5950
The goal of semi-supervised object detection is to learn a detection model using only a few labeled data and large amounts of unlabeled data, thereby reducing the cost of data labeling. Although a few studies have proposed various self-training-based methods or consistency regularization-based methods, they ignore the discrepancies among the detection results in the same image that occur during different training iterations. Additionally, the predicted detection results vary among different detection models. In this paper, we propose an interactive form of self-training using mean teachers for semi-supervised object detection. Specifically, to alleviate the instability among the detection results in different iterations, we propose using nonmaximum suppression to fuse the detection results from different iterations. Simultaneously, we use multiple detection heads that predict pseudo labels for each other to provide complementary information. Furthermore, to avoid different detection heads collapsing to each other, we use a mean teacher model instead of the original detection model to predict the pseudo labels. Thus, the object detection model can be trained on both labeled and unlabeled data. Extensive experimental results verify the effectiveness of our proposed method.
****
DeFLOCNet: Deep Image Editing via Flexible Low-Level Controls
Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, Jing Liao, Bin Jiang, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10765-10774
User-intended visual content fills the hole regions of an input image in the image editing scenario. The coarse lowlevel inputs, which typically consist of sparse sketch lines and color dots, convey user intentions for content creation (i.e., free-form editing). While existing methods combine an input image and these low-level controls for CNN inputs, the corresponding feature representations are not sufficient to convey user intentions, leading to unfaithfully generated content. In this paper, we propose DeFLOCNet which is based on a deep encoder-decoder CNN to retain the guidance of these controls in the deep feature representations. In each skip connection layer, we design a structure generation block. Instead of attaching low-level controls to an input image, we inject these controls directly into each structure generation block for sketch line refinement and color propagation in the CNN feature space. We then concatenate the modulated features with the original decoder features for structure generation. Meanwhile, DeFLOCNet involves another decoder branch for texture generation and detail enhancement. Both structures and textures are rendered in the decoder, leading to user-intended editing results. Experiments on benchmarks indicate that DeFLOCNet effectively transforms different user intentions to create visually pleasing content.
****
Vx2Text: End-to-End Learning of Video-Based Text Generation From Multimodal Inpu

ts
Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, Lorenzo Torre sani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2021, pp. 7005-7015
We present Vx2Text, a framework for text generation from multimodal inputs consi sting of video plus text, speech, or audio. In order to leverage transformer net works, which have been shown to be effective at modeling language, each modality is first converted into a set of language embeddings by a learnable tokenizer. This allows our approach to perform multimodal fusion in the language space, thu s eliminating the need for ad-hoc cross-modal fusion modules. To address the non -differentiability of tokenization on continuous inputs (e.g., video or audio), we utilize a relaxation scheme that enables end-to-end training. Furthermore, un like prior encoder-only models, our network includes an autoregressive decoder t o generate open-ended text from the multimodal embeddings fused by the language encoder. This renders our approach fully generative and makes it directly applic able to different "video+x to text" problems without the need to design speciali zed network heads for each task. The proposed framework is not only conceptually simple but also remarkably effective: experiments demonstrate that our approach based on a single architecture outperforms the state-of-the-art on three video-based text-generation tasks---captioning, question answering and audio-visual sc ene-aware dialog. Our code will be made publicly available.
****
KSM: Fast Multiple Task Adaption via Kernel-Wise Soft Mask Learning
Li Yang, Zhezhi He, Junshan Zhang, Deliang Fan; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13845-13853
Deep Neural Networks (DNN) could forget the knowledge about earlier tasks when l earning new tasks, and this is known as catastrophic forgetting. To learn new ta sk without forgetting, recently, the mask-based learning method (e.g. piggyback) is proposed to address these issues by learning only a binary element-wise mas k, while keeping the backbone model fixed. However, the binary mask has limited modeling capacity for new tasks. A more recent work proposes a compress-grow-bas ed method (CPG) to achieve better accuracy for new tasks by partially training b ackbone model, but with order-higher training cost, which makes it infeasible to be deployed into popular state-of-the-art edge-/mobile-learning. The primary go al of this work is to simultaneously achieve fast and high-accuracy multi-task a daption in a continual learning setting. Thus motivated, we propose a new traini ng method called Kernel-wise Soft Mask (KSM), which learns a kernel-wise hybrid binary and real-value soft mask for each task. Such a soft mask can be viewed as a superposition of a binary mask and a properly scaled real-value tensor, which offers a richer representation capability without low-level kernel support to m eet the objective of low hardware overhead. We validate KSM on multiple benchmar k datasets against recent state-of-the-art methods (e.g. Piggyback, Packnet, CPG , etc.), which shows good improvement in both accuracy and training cost.
****
Rich Context Aggregation With Reflection Prior for Glass Surface Detection
Jiaying Lin, Zebang He, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13415-13424
Glass surfaces appear everywhere. Their existence can however pose a serious pro blem to computer vision tasks. Recently, a method is proposed to detect glass su rfaces by learning multi-scale contextual information. However, as it is only ba sed on a general context integration operation and does not consider any specifi c glass surface properties, it gets confused when the images contain objects tha t are similar to glass surfaces and degenerates in challenging scenes with insuf ficient contexts. We observe that humans often rely on identifying reflections i n order to sense the existence of glass and on locating the boundary in order to determine the extent of the glass. Hence, we propose a model for glass surface detection, which consists of two novel modules: (1) a rich context aggregation m odule (RCAM) to extract multi-scale boundary features from rich context features for locating glass surface boundaries of different sizes and shapes, and (2) a reflection-based refinement module (RRM) to detect reflection and then incorpora

te it so as to differentiate glass regions from non-glass regions. In addition, we also propose a challenging dataset consisting of 4,012 glass images with annotations for glass surface detection. Our experiments demonstrate that the proposed model outperforms state-of-the-art methods from relevant fields.
****
Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization
Aysim Toker, Qunjie Zhou, Maxim Maximov, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6488-6497
The goal of cross-view image based geo-localization is to determine the location of a given street view image by matching it against a collection of geo-tagged satellite images. This task is notoriously challenging due to the drastic viewpoint and appearance differences between the two domains. We show that we can address this discrepancy explicitly by learning to synthesize realistic street views from satellite inputs. Following this observation, we propose a novel multi-task architecture in which image synthesis and retrieval are considered jointly. The rationale behind this is that we can bias our network to learn latent feature representations that are useful for retrieval if we utilize them to generate images across the two input domains. To the best of our knowledge, ours is the first approach that creates realistic street views from satellite images and localizes the corresponding query street view simultaneously in an end-to-end manner. In our experiments, we obtain state-of-the-art performance on the CVUSA and CVACT benchmarks. Finally, we show compelling qualitative results for satellite-to-street view synthesis.
****
AutoInt: Automatic Integration for Fast Neural Volume Rendering
David B. Lindell, Julien N. P. Martel, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14556-14565
Numerical integration is a foundational technique in scientific computing and is at the core of many computer vision applications. Among these applications, neural volume rendering has recently been proposed as a new paradigm for view synthesis, achieving photorealistic image quality. However, a fundamental obstacle to making these methods practical is the extreme computational and memory requirements caused by the required volume integrations along the rendered rays during training and inference. Millions of rays, each requiring hundreds of forward passes through a neural network are needed to approximate those integrations with Monte Carlo sampling. Here, we propose automatic integration, a new framework for learning efficient, closed-form solutions to integrals using coordinate-based neural networks. For training, we instantiate the computational graph corresponding to the derivative of the coordinate-based network. The graph is fitted to the signal to integrate. After optimization, we reassemble the graph to obtain a network that represents the antiderivative. By the fundamental theorem of calculus, this enables the calculation of any definite integral in two evaluations of the network. Applying this approach to neural rendering, we improve a tradeoff between rendering speed and image quality: improving render times by greater than 10x with a tradeoff of reduced image quality.
****
Pose-Guided Human Animation From a Single Image in the Wild
Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15039-15048
We present a new pose transfer method for synthesizing a human animation from a single image of a person controlled by a sequence of body poses. Existing pose transfer methods exhibit significant visual artifacts when applying to a novel scene, resulting in temporal inconsistency and failures in preserving the identity and textures of the person. To address these limitations, we design a compositional neural network that predicts the silhouette, garment labels, and textures. Each modular network is explicitly dedicated to a subtask that can be learned from the synthetic data. At the inference time, we utilize the trained network to

produce a unified representation of appearance and its labels in UV coordinates, which remain constant across poses. The unified representation provides incomplete yet strong guidance to generating the appearance in response to the pose change. We use the trained network to complete the appearance and render it with the background. With these strategies, we are able to synthesize human animations that can preserve the identity and appearance of the person in a temporally coherent way without any fine-tuning of the network on the testing scene. Experiments show that our method outperforms the state-of-the-arts in terms of synthesis quality, temporal coherence, and generalization ability.
****
Room-and-Object Aware Knowledge Reasoning for Remote Embodied Referring Expression
Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3064-3073
The Remote Embodied Referring Expression (REVERIE) is a recently raised task that requires an agent to navigate to and localise a referred remote object according to a high-level language instruction. Different from related VLN tasks, the key to REVERIE is to conduct goal-oriented exploration instead of strict instruction-following, due to the lack of step-by-step navigation guidance. In this paper, we propose a novel Cross-modality Knowledge Reasoning (CKR) model to address the unique challenges of this task. The CKR, based on a transformer-architecture, learns to generate scene memory tokens and utilise these informative history clues for exploration. Particularly, a Room-and-Object Aware Attention (ROAA) mechanism is devised to explicitly perceive the room- and object-type information from both linguistic and visual observations. Moreover, through incorporating commonsense knowledge, we propose a Knowledge-enabled Entity Relationship Reasoning (KERR) module to learn the internal-external correlations among room- and object-entities for agent to make proper action at each viewpoint. Evaluation on REVERIE benchmark demonstrates the superiority of the CKR model, which significantly boosts SPL and REVERIE-success rate by 64.67% and 46.05%, respectively. Code is available at: https://github.com/alloldman/CKR.
****
Equivariant Point Network for 3D Point Cloud Analysis
Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, Randall Hill; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14514-14523
Features that are equivariant to a larger group of symmetries have been shown to be more discriminative and powerful in recent studies. However, higher-order equivariant features often come with an exponentially-growing computational cost. Furthermore, it remains relatively less explored how rotation-equivariant features can be leveraged to tackle 3D shape alignment tasks. While many past approaches have been based on either non-equivariant or invariant descriptors to align 3D shapes, we argue that such tasks may benefit greatly from an equivariant framework. In this paper, we propose an effective and practical SE(3) (3D translation and rotation) equivariant network for point cloud analysis that addresses both problems. First, we present SE(3) separable point convolution, a novel framework that breaks down the 6D convolution into two separable convolutional operators alternatively performed in the 3D Euclidean and SO(3) spaces. This significantly reduces the computational cost without compromising the performance. Second, we introduce an attention layer to effectively harness the expressiveness of the equivariant features. While jointly trained with the network, the attention layer implicitly derives the intrinsic local frame in the feature space and generates attention vectors that can be integrated into different alignment tasks. We evaluate our approach through extensive studies and visual interpretations. The empirical results demonstrate that our proposed model outperforms strong baselines in a variety of benchmarks.
****
Learning Graph Embeddings for Compositional Zero-Shot Learning
Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, Zeynep Akata; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 953-962
In compositional zero-shot learning, the goal is to recognize unseen compositions (e.g. old dog) of observed visual primitives states (e.g. old, cute) and objects (e.g. car, dog)in the training set. This is challenging because the same state can for example alter the visual appearance of a dog drastically differently from a car. As a solution, we propose a novel graph formulation called Compositional Graph Embedding (CGE) that learns image features, compositional classifiers, and latent representations of visual primitives in an end-to-end manner. The key to our approach is exploit-ing the dependency between states, objects, and their compositions within a graph structure to enforce the relevant knowledge transfer from seen to unseen compositions. By learning a joint compatibility that encodes semantics between concepts, our model allows for generalization to un-seen compositions without relying on an external knowledgebase like WordNet. We show that in the challenging generalized compositional zero-shot setting our CGE significantly outperforms the state of the art on MIT-States and UT-Zappos. We also propose a new benchmark for this task based on the recent GQA dataset.
****
NeRD: Neural 3D Reflection Symmetry Detector
Yichao Zhou, Shichen Liu, Yi Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15940-15949
Recent advances have shown that symmetry, a structural prior that most objects exhibit, can support a variety of single-view 3D understanding tasks. However, detecting 3D symmetry from an image remains a challenging task. Previous works either assume the symmetry is given or detect the symmetry with a heuristic-based method. In this paper, we present NeRD, a Neural 3D Reflection Symmetry Detector, which combines the strength of learning-based recognition and geometry-based reconstruction to accurately recover the normal direction of objects' mirror planes. Specifically, we enumerate the symmetry planes with a coarse-to-fine strategy and find the best ones by building 3D cost volumes to examine the intra-image pixel correspondence from the symmetry. Our experiments show that the symmetry planes detected with our method are significantly more accurate than the planes from direct CNN regression on both synthetic and real datasets. More importantly, we also demonstrate that the detected symmetry can be used to improve the performance of downstream tasks such as pose estimation and depth map regression by a wide margin over existing methods. The code of this paper has been made public at https://github.com/zhou13/nerd.
****
Checkerboard Context Model for Efficient Learned Image Compression
Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, Hongwei Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14771-14780
For learned image compression, the autoregressive context model is proved effective in improving the rate-distortion (RD) performance. Because it helps remove spatial redundancies among latent representations. However, the decoding process must be done in a strict scan order, which breaks the parallelization. We propose a parallelizable checkerboard context model (CCM) to solve the problem. Our two-pass checkerboard context calculation eliminates such limitations on spatial locations by re-organizing the decoding order. Speeding up the decoding process more than 40 times in our experiments, it achieves significantly improved computational efficiency with almost the same rate-distortion performance. To the best of our knowledge, this is the first exploration on parallelization-friendly spatial context model for learned image compression.
****
Zero-Shot Adversarial Quantization
Yuang Liu, Wei Zhang, Jun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1512-1521
Model quantization is a promising approach to compress deep neural networks and accelerate inference, making it possible to be deployed on mobile and edge devices. To retain the high performance of full-precision models, most existing quant

ization methods focus on fine-tuning quantized model by assuming training datase
ts are accessible. However, this assumption sometimes is not satisfied in real s
ituations due to data privacy and security issues, thereby making these quantiza
tion methods not applicable. To achieve zero-short model quantization without ac
cessing training data, a tiny number of quantization methods adopt either post-t
raining quantization or batch normalization statistics-guided data generation fo
r fine-tuning. However, both of them inevitably suffer from low performance, sin
ce the former is a little too empirical and lacks training support for ultra-low
 precision quantization, while the latter could not fully restore the peculiarit
ies of original data and is often low efficient for diverse data generation. To
address the above issues, we propose a zero-shot adversarial quantization (ZAQ)
framework, facilitating effective discrepancy estimation and knowledge transfer
from a full-precision model to its quantized model. This is achieved by a novel
two-level discrepancy modeling to drive a generator to synthesize informative an
d diverse data examples to optimize the quantized model in an adversarial learni
ng fashion. We conduct extensive experiments on three fundamental vision tasks,
demonstrating the superiority of ZAQ over the strong zero-shot baselines and val
idating the effectiveness of its main components.
****
Group Whitening: Balancing Learning Efficiency and Representational Capacity
Lei Huang, Yi Zhou, Li Liu, Fan Zhu, Ling Shao; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9512-9521
Batch normalization (BN) is an important technique commonly incorporated into de
ep learning models to perform standardization within mini-batches. The merits of
 BN in improving a model's learning efficiency can be further amplified by apply
ing whitening, while its drawbacks in estimating population statistics for infer
ence can be avoided through group normalization (GN). This paper proposes group
whitening (GW), which exploits the advantages of the whitening operation and avo
ids the disadvantages of normalization within mini-batches. In addition, we anal
yze the constraints imposed on features by normalization, and show how the batch
 size (group number) affects the performance of batch (group) normalized network
s, from the perspective of model's representational capacity. This analysis prov
ides theoretical guidance for applying GW in practice. Finally, we apply the pro
posed GW to ResNet and ResNeXt architectures and conduct experiments on the Imag
eNet and COCO benchmarks. Results show that GW consistently improves the perform
ance of different architectures, with absolute gains of 1.02% 1.49% in top-1 ac
curacy on ImageNet and 1.82% 3.21% in bounding box AP on COCO.
****
Adversarial Robustness Under Long-Tailed Distribution
Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, Dahua Lin; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8659-
8668
Adversarial robustness has attracted extensive studies recently by revealing the
 vulnerability and intrinsic characteristics of deep networks. However, existing
 works on adversarial robustness mainly focus on balanced datasets, while real-w
orld data usually exhibits a long-tailed distribution. To push adversarial robus
tness towards more realistic scenarios, in this work we investigate the adversar
ial vulnerability as well as defense under long-tailed distributions. In particu
lar, we first reveal the negative impacts induced by imbalanced data on both rec
ognition performance and adversarial robustness, uncovering the intrinsic challe
nges of this problem. We then perform a systematic study on existing long-tailed
 recognition methods in conjunction with the adversarial training framework. Sev
eral valuable observations are obtained: 1) natural accuracy is relatively easy
to improve, 2) fake gain of robust accuracy exists under unreliable evaluation,
and 3) boundary error limits the promotion of robustness. Inspired by these obse
rvations, we propose a clean yet effective framework, RoBal, which consists of t
wo dedicated modules, a scale-invariant classifier and data re-balancing via bot
h margin engineering at training stage and boundary adjustment during inference.
 Extensive experiments demonstrate the superiority of our approach over other st
ate-of-the-art defense methods. To our best knowledge, we are the first to tackl

e adversarial robustness under long-tailed distributions, which we believe would be a significant step towards real-world robustness. Our code is available at: https://github.com/wutong16/Adversarial_Long-Tail.
****

HyperSeg: Patch-Wise Hypernetwork for Real-Time Semantic Segmentation
Yuval Nirkin, Lior Wolf, Tal Hassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4061-4070
We present a novel, real-time, semantic segmentation network in which the encoder both encodes and generates the parameters (weights) of the decoder. Furthermore, to allow maximal adaptivity, the weights at each decoder block vary spatially. For this purpose, we design a new type of hypernetwork, composed of a nested U-Net for drawing higher level context features, a multi-headed weight generating module which generates the weights of each block in the decoder immediately before they are consumed, for efficient memory utilization, and a primary network that is composed of novel dynamic patch-wise convolutions. Despite the usage of less-conventional blocks, our architecture obtains real-time performance. In terms of the runtime vs. accuracy trade-off, we surpass state of the art (SotA) results on popular semantic segmentation benchmarks: PASCAL VOC 2012 (val. set) and real-time semantic segmentation on Cityscapes, and CamVid. The code is available: https://nirkin.com/hyperseg.
****

Augmentation Strategies for Learning With Noisy Labels
Kento Nishi, Yi Ding, Alex Rich, Tobias Hollerer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8022-8031
Imperfect labels are ubiquitous in real-world datasets. Several recent successful methods for training deep neural networks (DNNs) robust to label noise have used two primary techniques: filtering samples based on loss during a warm-up phase to curate an initial set of cleanly labeled samples, and using the output of a network as a pseudo-label for subsequent loss calculations. In this paper, we evaluate different augmentation strategies for algorithms tackling the ""learning with noisy labels"" problem. We propose and examine multiple augmentation strategies and evaluate them using synthetic datasets based on CIFAR-10 and CIFAR-100, as well as on the real-world dataset Clothing1M. Due to several commonalities in these algorithms, we find that using one set of augmentations for loss modeling tasks and another set for learning is the most effective, improving results on the state-of-the-art and other previous methods. Furthermore, we find that applying augmentation during the warm-up period can negatively impact the loss convergence behavior of correctly versus incorrectly labeled samples. We introduce this augmentation strategy to the state-of-the-art technique and demonstrate that we can improve performance across all evaluated noise levels. In particular, we improve accuracy on the CIFAR-10 benchmark at 90% symmetric noise by more than 15% in absolute accuracy, and we also improve performance on the Clothing1M dataset.
****

AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching
Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, Jianping Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10328-10337
Recently, records on stereo matching benchmarks are constantly broken by end-to-end disparity networks. However, the domain adaptation ability of these deep models is quite poor. Addressing such problem, we present a novel domain-adaptive pipeline called AdaStereo that aims to align multi-level representations for deep stereo matching networks. Compared to previous methods for adaptive stereo matching, our AdaStereo realizes a more standard, complete and effective domain adaptation pipeline. Firstly, we propose a non-adversarial progressive color transfer algorithm for input image-level alignment. Secondly, we design an efficient parameter-free cost normalization layer for internal feature-level alignment. Lastly, a highly related auxiliary task, self-supervised occlusion-aware reconstruction is presented to narrow down the gaps in output space. Our AdaStereo models achieve state-of-the-art cross-domain performance on multiple stereo benchmarks,

including KITTI, Middlebury, ETH3D, and DrivingStereo, even outperforming disparity networks finetuned with target-domain ground-truths.
****
ClassSR: A General Framework to Accelerate Super-Resolution Networks by Data Characteristic
Xiangtao Kong, Hengyuan Zhao, Yu Qiao, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12016-12025
We aim at accelerating super-resolution (SR) networks on large images (2K-8K). The large images are usually decomposed into small sub-images in practical usages. Based on this processing, we found that different image regions have different restoration difficulties and can be processed by networks with different capacities. Intuitively, smooth areas are easier to super-solve than complex textures. To utilize this property, we can adopt appropriate SR networks to process different sub-images after the decomposition. On this basis, we propose a new solution pipeline -- ClassSR that combines classification and SR in a unified framework. In particular, it first uses a Class-Module to classify the sub-images into different classes according to restoration difficulties, then applies an SR-Module to perform SR for different classes. The Class-Module is a conventional classification network, while the SR-Module is a network container that consists of the to-be-accelerated SR network and its simplified versions. We further introduce a new classification method with two losses -- Class-Loss and Average-Loss to produce the classification results. After joint training, a majority of sub-images will pass through smaller networks, thus the computational cost can be significantly reduced. Experiments show that our ClassSR can help most existing methods (e.g., FSRCNN, CARN, SRResNet, RCAN) save up to 50% FLOPs on DIV8K datasets. This general framework can also be applied in other low-level vision tasks.
****
Partition-Guided GANs
Mohammadreza Armandpour, Ali Sadeghian, Chunyuan Li, Mingyuan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5099-5109
Despite the success of Generative Adversarial Networks (GANs), their training suffers from several well-known problems, including mode collapse and difficulties learning a disconnected set of manifolds. In this paper, we break down the challenging task of learning complex high dimensional distributions, supporting diverse data samples, to simpler sub-tasks. Our solution relies on designing a partitioner that breaks the space into smaller regions, each having a simpler distribution, and training a different generator for each partition. This is done in an unsupervised manner without requiring any labels. We formulate two desired criteria for the space partitioner that aid the training of our mixture of generators: 1) to produce connected partitions and 2) provide a proxy of distance between partitions and data samples, along with a direction for reducing that distance. These criteria are developed to avoid producing samples from places with non-existent data density, and also facilitate training by providing additional direction to the generators. We develop theoretical constraints for a space partitioner to satisfy the above criteria. Guided by our theoretical analysis, we design an effective neural architecture for the space partitioner that empirically assures these conditions. Experimental results on various standard benchmarks show that the proposed unsupervised model outperforms several recent methods.
****
GATSBI: Generative Agent-Centric Spatio-Temporal Object Interaction
Cheol-Hui Min, Jinseok Bae, Junho Lee, Young Min Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3074-3083
We present GATSBI, a generative model that can transform a sequence of raw observations into a structured latent representation that fully captures the spatio-temporal context of the agent's actions. In vision-based decision-making scenarios, an agent faces complex high-dimensional observations where multiple entities interact with each other. The agent requires a good scene representation of the

visual observation that discerns essential components that consistently propagates along the time horizon. Our method, GATSBI, utilizes unsupervised scene representation learning to successfully separate an active agent, static background, and passive objects. GATSBI then models the interactions reflecting the causal relationships among decomposed entities and predicts physically plausible future states. Our model generalizes to a variety of environments where different types of robots and objects dynamically interact with each other. GATSBI achieves superior performance on scene decompo-sition and video prediction compared to its state-of-the-artcounterparts, and can be readily applied to sequential deci-sion making of an intelligent agent.
****
Privacy-Preserving Collaborative Learning With Automatic Transformation Search
Wei Gao, Shangwei Guo, Tianwei Zhang, Han Qiu, Yonggang Wen, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 114-123
Collaborative learning has gained great popularity due to its benefit of data privacy protection: participants can jointly train a Deep Learning model without sharing their training sets. However, recent works discovered that an adversary can fully recover the sensitive training samples from the shared gradients. Such reconstruction attacks pose severe threats to collaborative learning. Hence, effective mitigation solutions are urgently desired. In this paper, we propose to leverage data augmentation to defeat reconstruction attacks: by preprocessing sensitive images with carefully-selected transformation policies, it becomes infeasible for the adversary to extract any useful information from the corresponding gradients. We design a novel search method to automatically discover qualified policies. We adopt two new metrics to quantify the impacts of transformations on data privacy and model usability, which can significantly accelerate the search speed. Comprehensive evaluations demonstrate that the policies discovered by our method can defeat existing reconstruction attacks in collaborative learning, with high efficiency and negligible impact on the model performance.
****
Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval
Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, Zheng Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2215-2224
Given an untrimmed video and a query sentence, cross-modal video moment retrieval aims to rank a video moment from pre-segmented video moment candidates that best matches the query sentence. Pioneering work typically learns the representations of the textual and visual content separately and then obtains the interactions or alignments between different modalities. However, the task of cross-modal video moment retrieval is not yet thoroughly addressed as it needs to further identify the fine-grained differences of video moment candidates with high repeatability and similarity. Moveover, the relation among objects in both video and query sentence is intuitive and efficient for understanding semantics but is rarely considered. Toward this end, we contribute a multi-modal relational graph to capture the interactions among objects from the visual and textual content to identify the differences among similar video moment candidates. Specifically, we first introduce a visual relational graph and a textual relational graph to form relation-aware representations via message propagation. Thereafter, a multi-task pre-training is designed to capture domain-specific knowledge about objects and relations, enhancing the structured visual representation after explicitly defined relation. Finally, the graph matching and boundary regression are employed to perform the cross-modal retrieval. We conduct extensive experiments on two datasets about daily activities and cooking activities, demonstrating significant improvements over state-of-the-art solutions.
****
Point Cloud Instance Segmentation Using Probabilistic Embeddings
Biao Zhang, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8883-8892
In this paper, we propose a new framework for point cloud instance segmentation.

Our framework has two steps: an embedding step and a clustering step. In the em
bedding step, our main contribution is to propose a probabilistic embedding spac
e for point cloud embedding. Specifically, each point is represented as a tri-va
riate normal distribution. In the clustering step, we propose a novel loss funct
ion, which benefits both the semantic segmentation and the clustering. Our exper
imental results show important improvements to the SOTA, i.e., 3.1% increased av
erage per-category mAP on the PartNet dataset.
****
pixelNeRF: Neural Radiance Fields From One or Few Images
Alex Yu, Vickie Ye, Matthew Tancik, Angjoo Kanazawa; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4578-45
87
We propose pixelNeRF, a learning framework that predicts a continuous neural sce
ne representation conditioned on one or few input images. The existing approach
for constructing neural radiance fields (NeRFs) involves optimizing the represen
tation to every scene independently, requiring many calibrated views and signifi
cant compute time. We take a step towards resolving these shortcomings by introd
ucing an architecture that conditions a NeRF on image inputs in a fully convolut
ional manner. This allows the network to be trained across multiple scenes to le
arn a scene prior, allowing it to perform novel view synthesis in a feed-forward
 manner from a sparse set of views (as few as one). Leveraging the volume render
ing approach of NeRF, our model can be trained directly from images with no expl
icit 3D supervision. We conduct extensive experiments on ShapeNet benchmarks for
 single image novel view synthesis tasks under category specific and category ag
nostic settings. We further demonstrate the flexibility of pixelNeRF by demonstr
ating it on multi-object ShapeNet scenes as well as real scenes from the DTU dat
aset. In all cases, pixelNeRF outperforms current state-of-the-art baselines for
 novel view synthesis and single image 3D reconstruction.
****
Navigating the GAN Parameter Space for Semantic Image Editing
Anton Cherepkov, Andrey Voynov, Artem Babenko; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3671-3680
Generative Adversarial Networks (GANs) are currently an indispensable tool for v
isual editing, being a standard component of image-to-image translation and imag
e restoration pipelines. Furthermore, GANs are especially useful for controllabl
e generation since their latent spaces contain a wide range of interpretable dir
ections, well suited for semantic editing operations. By gradually changing late
nt codes along these directions, one can produce impressive visual effects, unat
tainable without GANs. In this paper, we significantly expand the range of visua
l effects achievable with the state-of-the-art models, like StyleGAN2. In contra
st to existing works, which mostly operate by latent codes, we discover interpr
etable directions in the space of the generator parameters. By several simple me
thods, we explore this space and demonstrate that it also contains a plethora of
 interpretable directions, which are an excellent source of non-trivial semantic
 manipulations. The discovered manipulations cannot be achieved by transforming
the latent codes and can be used to edit both synthetic and real images. We rele
ase our code and models and hope they will serve as a handy tool for further eff
orts on GAN-based image editing.
****
Large-Capacity Image Steganography Based on Invertible Neural Networks
Shao-Ping Lu, Rong Wang, Tao Zhong, Paul L. Rosin; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10816-108
25
Many attempts have been made to hide information in images, where the main chall
enge is how to increase the payload capacity without the container image being d
etected as containing a message. In this paper, we propose a large-capacity Inve
rtible Steganography Network (ISN) for image steganography. We take steganograph
y and the recovery of hidden images as a pair of inverse problems on image domai
n transformation, and then introduce the forward and backward propagation operat
ions of a single invertible network to leverage the image embedding and extracti

ng problems. Sharing all parameters of our single ISN architecture enables us to efficiently generate both the container image and the revealed hidden image(s) with high quality. Moreover, in our architecture the capacity of image steganography is significantly improved by naturally increasing the number of channels of the hidden image branch. Comprehensive experiments demonstrate that with this significant improvement of the steganography capacity, our ISN achieves state-of-the-art in both visual and quantitative comparisons.
****
Exploiting Edge-Oriented Reasoning for 3D Point-Based Scene Graph Analysis
Chaoyi Zhang, Jianhui Yu, Yang Song, Weidong Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9705-9715
Scene understanding is a critical problem in computer vision. In this paper, we propose a 3D point-based scene graph generation (SGGpoint) framework to effectively bridge perception and reasoning to achieve scene understanding via three sequential stages, namely scene graph construction, reasoning, and inference. Within the reasoning stage, an EDGE-oriented Graph Convolutional Network (EdgeGCN) is created to exploit multi-dimensional edge features for explicit relationship modeling, together with the exploration of two associated twinning interaction mechanisms between nodes and edges for the independent evolution of scene graph representations. Overall, our integrated SGGpoint framework is established to seek and infer scene structures of interest from both real-world and synthetic 3D point-based scenes. Our experimental results show promising edge-oriented reasoning effects on scene graph generation studies. We also demonstrate our method advantage on several traditional graph representation learning benchmark datasets, including the node-wise classification on citation networks and whole-graph recognition problems for molecular analysis.
****
CoLA: Weakly-Supervised Temporal Action Localization With Snippet Contrastive Learning
Can Zhang, Meng Cao, Dongming Yang, Jie Chen, Yuexian Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16010-16019
Weakly-supervised temporal action localization (WS-TAL) aims to localize actions in untrimmed videos with only video-level labels. Most existing models follow the "localization by classification" procedure: locate temporal regions contributing most to the video-level classification. Generally, they process each snippet (or frame) individually and thus overlook the fruitful temporal context relation. Here arises the single snippet cheating issue: "hard" snippets are too vague to be classified. In this paper, we argue that learning by comparing helps identify these hard snippets and we propose to utilize snippet Contrastive learning to Localize Actions, CoLA for short. Specifically, we propose a Snippet Contrast (SniCo) Loss to refine the hard snippet representation in feature space, which guides the network to perceive precise temporal boundaries and avoid the temporal interval interruption. Besides, since it is infeasible to access frame-level annotations, we introduce a Hard Snippet Mining algorithm to locate the potential hard snippets. Substantial analyses verify that this mining strategy efficaciously captures the hard snippets and SniCo Loss leads to more informative feature representation. Extensive experiments show that CoLA achieves state-of-the-art results on THUMOS'14 and ActivityNet v1.2 datasets.
****
MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition
Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, Xinjing Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5212-5221
Real-world training data usually exhibits long-tailed distribution, where several majority classes have a significantly larger number of samples than the remaining minority classes. This imbalance degrades the performance of typical supervised learning algorithms designed for balanced training sets. In this paper, we address this issue by augmenting minority classes with a recently proposed implicit semantic data augmentation (ISDA) algorithm, which produces diversified augme

nted samples by translating deep features along many semantically meaningful dir
ections. Importantly, given that ISDA estimates the class-conditional statistics
 to obtain semantic directions, we find it ineffective to do this on minority cl
asses due to the insufficient training data. To this end, we propose a novel app
roach to learn transformed semantic directions with meta-learning automatically.
 In specific, the augmentation strategy during training is dynamically optimized
, aiming to minimize the loss on a small balanced validation set, which is appro
ximated via a meta update step. Extensive empirical results on CIFAR-LT-10/100,
ImageNet-LT, and iNaturalist 2017/2018 validate the effectiveness of our method.
****

Limitations of Post-Hoc Feature Alignment for Robustness
Collin Burns, Jacob Steinhardt; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 2525-2533
Feature alignment is an approach to improving robustness to distribution shift t
hat matches the distribution of feature activations between the training distrib
ution and test distribution. A particularly simple but effective approach to fea
ture alignment involves aligning the batch normalization statistics between the
two distributions in a trained neural network. This technique has received renew
ed interest lately because of its impressive performance on robustness benchmark
s. However, when and why this method works is not well understood. We investigat
e the approach in more detail and identify several limitations. We show that it
only significantly helps with a narrow set of distribution shifts and we identif
y several settings in which it even degrades performance. We also explain why th
ese limitations arise by pinpointing why this approach can be so effective in th
e first place. Our findings call into question the utility of this approach and
Unsupervised Domain Adaptation more broadly for improving robustness in practice
.
****

Every Annotation Counts: Multi-Label Deep Supervision for Medical Image Segmenta
tion
Simon Reiss, Constantin Seibold, Alexander Freytag, Erik Rodner, Rainer Stiefelh
agen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2021, pp. 9532-9542
Pixel-wise segmentation is one of the most data and annotation hungry tasks in o
ur field. Providing representative and accurate annotations is often mission-cri
tical especially for challenging medical applications. In this paper, we propose
 a semi-weakly supervised segmentation algorithm to overcome this barrier. Our a
pproach is based on a new formulation of deep supervision and student-teacher mo
del and allows for easy integration of different supervision signals. In contras
t to previous work, we show that care has to be taken how deep supervision is in
tegrated in lower layers and we present multi-label deep supervision as the most
 important secret ingredient for success. With our novel training regime for seg
mentation that flexibly makes use of images that are either fully labeled, marke
d with bounding boxes, just global labels, or not at all, we are able to cut the
 requirement for expensive labels by 94.22% - narrowing the gap to the best full
y supervised baseline to only 5% mean IoU. Our approach is validated by extensiv
e experiments on retinal fluid segmentation and we provide an in-depth analysis
of the anticipated effect each annotation type can have in boosting segmentation
 performance.
****

Roses Are Red, Violets Are Blue... but Should VQA Expect Them To?
Corentin Kervadec, Grigory Antipov, Moez Baccouche, Christian Wolf; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
21, pp. 2776-2785
Models for Visual Question Answering (VQA) are notorious for their tendency to r
ely on dataset biases, as the large and unbalanced diversity of questions and co
ncepts involved and tends to prevent models from learning to ""reason"", leading
 them to perform ""educated guesses"" instead. In this paper, we claim that the
standard evaluation metric, which consists in measuring the overall in-domain ac
curacy, is misleading. Since questions and concepts are unbalanced, this tends t

o favor models which exploit subtle training set statistics. Alternatively, naively introducing artificial distribution shifts between train and test splits is also not completely satisfying. First, the shifts do not reflect real-world tendencies, resulting in unsuitable models; second, since the shifts are handcrafted, trained models are specifically designed for this particular setting, and do not generalize to other configurations. We propose the GQA-OOD benchmark designed to overcome these concerns: we measure and compare accuracy over both rare and frequent question-answer pairs, and argue that the former is better suited to the evaluation of reasoning abilities, which we experimentally validate with models trained to more or less exploit biases. In a large-scale study involving 7 VQA models and 3 bias reduction techniques, we also experimentally demonstrate that these models fail to address questions involving infrequent concepts and provide recommendations for future directions of research.
****
FAPIS: A Few-Shot Anchor-Free Part-Based Instance Segmenter
Khoi Nguyen, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11099-11108
This paper is about few-shot instance segmentation, where training and test image sets do not share the same object classes. We specify and evaluate a new few-shot anchor-free part-based instance segmenter (FAPIS). Our key novelty is in explicit modeling of latent object parts shared across training object classes, which is expected to facilitate our few-shot learning on new classes in testing. We specify a new anchor-free object detector aimed at scoring and regressing locations of foreground bounding boxes, as well as estimating relative importance of latent parts within each box. Also, we specify a new network for delineating and weighting latent parts for the final instance segmentation within every detected bounding box. Our evaluation on the benchmark COCO-20i dataset demonstrates that we significantly outperform the state of the art.
****
Disentangling Label Distribution for Long-Tailed Visual Recognition
Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, Buru Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6626-6636
The current evaluation protocol of long-tailed visual recognition trains the classification model on the long-tailed source label distribution and evaluates its performance on the uniform target label distribution. Such protocol has questionable practicality since the target may also be long-tailed. Therefore, we formulate long-tailed visual recognition as a label shift problem where the target and source label distributions are different. One of the significant hurdles in dealing with the label shift problem is the entanglement between the source label distribution and the model prediction. In this paper, we focus on disentangling the source label distribution from the model prediction. We first introduce a simple but overlooked baseline method that matches the target label distribution by post-processing the model prediction trained by the cross-entropy loss and the Softmax function. Although this method surpasses state-of-the-art methods on benchmark datasets, it can be further improved by directly disentangling the source label distribution from the model prediction in the training phase. Thus, we propose a novel method, LAbel distribution DisEntangling (LADE) loss based on the optimal bound of Donsker-Varadhan representation. LADE achieves state-of-the-art performance on benchmark datasets such as CIFAR-100-LT, Places-LT, ImageNet-LT, and iNaturalist 2018. Moreover, LADE outperforms existing methods on various shifted target label distributions, showing the general adaptability of our proposed method.
****
Gradient Forward-Propagation for Large-Scale Temporal Video Modelling
Mateusz Malinowski, Dimitrios Vytiniotis, Grzegorz Swirszcz, Viorica Patraucean, Joao Carreira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9249-9259
How can neural networks be trained on large-volume temporal data efficiently? To compute the gradients required to update parameters, backpropagation blocks com

putations until the forward and backward passes are completed. For temporal signals, this introduces high latency and hinders real-time learning. It also creates a coupling between consecutive layers, which limits model parallelism and increases memory consumption. In this paper, we build upon Sideways, which avoids blocking by propagating approximate gradients forward in time, by proposing mechanisms for temporal integration of information based on different variants of skip connections. We also show how to decouple computation and delegate individual neural modules to different devices, allowing distributed and parallel training. The proposed Skip-sideways achieves low latency training, model parallelism, and, importantly, is capable of extracting temporal features, leading to more stable training and improved performance on real-world video datasets such as HMDB51, UCF101, and the large-scale Kinetics600. Finally, we also show that models trained with Skip-sideways generate better future frames than Sideways models, and hence they can better utilize motion cues.
****
Learning a Non-Blind Deblurring Network for Night Blurry Images
Liang Chen, Jiawei Zhang, Jinshan Pan, Songnan Lin, Faming Fang, Jimmy S. Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10542-10550
Deblurring night blurry images is difficult, because the common-used blur model based on the linear convolution operation does not hold in this situation due to the influence of saturated pixels. In this paper, we propose a non-blind deblurring network (NBDN) to restore night blurry images. To mitigate the side effects brought by the pixels that violate the blur model, we develop a confidence estimation unit (CEU) to estimate a map which ensures smaller contributions of these pixels to the deconvolution steps that are further optimized by the conjugate gradient (CG) method. Moreover, unlike the existing methods using manually tuned hyper-parameters in their frameworks, we propose a hyper-parameter estimation unit (HPEU) to adaptively estimate hyper-parameters for better image restoration. The experimental results demonstrate that the proposed network performs favorably against state-of-the-art algorithms both quantitatively and qualitatively.
****
Differentiable Diffusion for Dense Depth Estimation From Multi-View Images
Numair Khan, Min H. Kim, James Tompkin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8912-8921
We present a method to estimate dense depth by optimizing a sparse set of points such that their diffusion into a depth map minimizes a multi-view reprojection error from RGB supervision. We optimize point positions, depths, and weights with respect to the loss by differential splatting that models points as Gaussians with analytic transmittance. Further, we develop an efficient optimization routine that can simultaneously optimize the 50k+ points required for complex scene reconstruction. We validate our routine using ground truth data and show high reconstruction quality. Then, we apply this to light field and wider baseline images via self supervision, and show improvements in both average and outlier error for depth maps diffused from inaccurate sparse points. Finally, we compare qualitative and quantitative results to image processing and deep learning methods.
****
Deep Compositional Metric Learning
Wenzhao Zheng, Chengkun Wang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9320-9329
In this paper, we propose a deep compositional metric learning (DCML) framework for effective and generalizable similarity measurement between images. Conventional deep metric learning methods minimize a discriminative loss to enlarge interclass distances while suppressing intraclass variations, which might lead to inferior generalization performance since samples even from the same class may present diverse characteristics. This motivates the adoption of the ensemble technique to learn a number of sub-embeddings using different and diverse subtasks. However, most subtasks impose weaker or contradictory constraints, which essentially sacrifices the discrimination ability of each sub-embedding to improve the generalization ability of their combination. To achieve a better generalization abi

lity without compromising, we propose to separate the sub-embeddings from direct supervisions from the subtasks and apply the losses on different composites of the sub-embeddings. We employ a set of learnable compositors to combine the sub-embeddings and use a self-reinforced loss to train the compositors, which serve as relays to distribute the diverse training signals to avoid destroying the dis crimination ability. Experimental results on the CUB-200-2011, Cars196, and Stan ford Online Products datasets demonstrate the superior performance of our framew ork.
****
Representing Videos As Discriminative Sub-Graphs for Action Recognition
Dong Li, Zhaofan Qiu, Yingwei Pan, Ting Yao, Houqiang Li, Tao Mei; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 1, pp. 3310-3319
Human actions are typically of combinatorial structures or patterns, i.e., subje cts, objects, plus spatio-temporal interactions in between. Discovering such str uctures is therefore a rewarding way to reason about the dynamics of interaction s and recognize the actions. In this paper, we introduce a new design of sub-gra phs to represent and encode the discriminative patterns of each action in the vi deos. Specifically, we present MUlti-scale Sub-graph LEarning (MUSLE) framework that novelly builds space-time graphs and clusters the graphs into compact sub-g raphs on each scale with respect to the number of nodes. Technically, MUSLE prod uces 3D bounding boxes, i.e., tubelets, in each video clip, as graph nodes and t akes dense connectivity as graph edges between tubelets. For each action categor y, we execute online clustering to decompose the graph into sub-graphs on each s cale through learning Gaussian Mixture Layer and select the discriminative sub-g raphs as action prototypes for recognition. Extensive experiments are conducted on both Something-Something V1 & V2 and Kinetics-400 datasets, and superior resu lts are reported when comparing to state-of-the-art methods. More remarkably, ou r MUSLE achieves to-date the best reported accuracy of 65.0% on Something-Someth ing V2 validation set.
****
AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training
Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, Cristian Sminch isescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2021, pp. 9919-9928
I went to the gym today, but how well did I do? And where should I improve? Ah, my back hurts slightly... User engagement can be sustained and injuries avoided by being able to reconstruct 3d human pose and motion, relate it to good trainin g practices, identify errors, and provide early, real-time feedback. In this pap er we introduce the first automatic system, AIFit, that performs 3d human sensin g for fitness training. The system can be used at home, outdoors, or at the gym. AIFit is able to reconstruct 3d human pose and motion, reliably segment exercis e repetitions, and identify in real-time the deviations between standards learnt from trainers, and the execution of a trainee. As a result, localized, quantita tive feedback for correct execution of exercises, reduced risk of injury, and co ntinuous improvement is possible. To support research and evaluation, we introdu ce the first large scale dataset, Fit3D, containing over 3 million images and co rresponding 3d human shape and motion capture ground truth configurations, with over 37 repeated exercises, covering all the major muscle groups, performed by i nstructors and trainees. Our statistical coach is governed by a global parameter that captures how critical it should be of a trainee's performance. This is an important aspect that helps adapt to a student's level of fitness (i.e. beginner vs. advanced vs. expert), or to the expected accuracy of a 3d pose reconstructi on method. We show that, for different values of the global parameter, our feedb ack system based on 3d pose estimates achieves good accuracy compared to the one based on ground-truth motion capture. Our statistical coach offers feedback in natural language, and with spatio-temporal visual grounding.
****
Synthesizing Long-Term 3D Human Motion and Interaction in 3D Scenes
Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, Xiaolong Wang; Proceedings of th

e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9401-9411
Synthesizing 3D human motion plays an important role in many graphics applications as well as understanding human activity. While many efforts have been made on generating realistic and natural human motion, most approaches neglect the importance of modeling human-scene interactions and affordances. On the other hand, affordance reasoning (e.g., standing on the floor or sitting on the chair) has mainly been studied with static human pose and gestures, and it has rarely been addressed with human motion. In this paper, we propose to bridge human motion synthesis and scene affordance reasoning. We present a hierarchical generative framework which synthesizes long-term 3D human motion conditioning on the 3D scene structure. We also further enforce multiple geometry constraints between the human mesh and scene point clouds via optimization to improve realistic synthesis. Our experiments show significant improvements over previous approaches on generating natural and physically plausible human motion in a scene.
****
How Well Do Self-Supervised Models Transfer?
Linus Ericsson, Henry Gouk, Timothy M. Hospedales; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5414-5423
Self-supervised visual representation learning has seen huge progress recently, but no large scale evaluation has compared the many models now available. We evaluate the transfer performance of 13 top self-supervised models on 40 downstream tasks, including many-shot and few-shot recognition, object detection, and dense prediction. We compare their performance to a supervised baseline and show that on most tasks the best self-supervised models outperform supervision, confirming the recently observed trend in the literature. We find ImageNet Top-1 accuracy to be highly correlated with transfer to many-shot recognition, but increasingly less so for few-shot, object detection and dense prediction. No single self-supervised method dominates overall, suggesting that universal pre-training is still unsolved. Our analysis of features suggests that top self-supervised learners fail to preserve colour information as well as supervised alternatives, but tend to induce better classifier calibration, and less attentive overfitting than supervised learners.
****
Understanding Object Dynamics for Interactive Image-to-Video Synthesis
Andreas Blattmann, Timo Milbich, Michael Dorkenwald, Bjorn Ommer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5171-5181
What would be the effect of locally poking a static scene? We present an approach that learns naturally-looking global articulations caused by a local manipulation at a pixel level. Training requires only videos of moving objects but no information of the underlying manipulation of the physical scene. Our generative model learns to infer natural object dynamics as a response to user interaction and learns about the interrelations between different object body regions. Given a static image of an object and a local poking of a pixel, the approach then predicts how the object would deform over time. In contrast to existing work on video prediction, we do not synthesize arbitrary realistic videos but enable local interactive control of the deformation. Our model is not restricted to particular object categories and can transfer dynamics onto novel unseen object instances. Extensive experiments on diverse objects demonstrate the effectiveness of our approach compared to common video prediction frameworks. Project page is available at https://bit.ly/3cxfA2L.
****
Pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis
Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5799-5809
We have witnessed rapid progress on 3D-aware image synthesis, leveraging recent advances in generative visual models and neural rendering. Existing approaches h

owever fall short in two ways: first, they may lack an underlying 3D representation or rely on view-inconsistent rendering, hence synthesizing images that are not multi-view consistent; second, they often depend upon representation network architectures that are not expressive enough, and their results thus lack in image quality. We propose a novel generative model, named Periodic Implicit Generative Adversarial Networks (p-GAN or pi-GAN), for high-quality 3D-aware image synthesis. p-GAN leverages neural representations with periodic activation functions and volumetric rendering to represent scenes as view-consistent radiance fields. The proposed approach obtains state-of-the-art results for 3D-aware image synthesis with multiple real and synthetic datasets.
****
Diverse Branch Block: Building a Convolution as an Inception-Like Unit
Xiaohan Ding, Xiangyu Zhang, Jungong Han, Guiguang Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10886-10895
We propose a universal building block of Convolutional Neural Network (ConvNet) to improve the performance without any inference-time costs. The block is named Diverse Branch Block (DBB), which enhances the representational capacity of a single convolution by combining diverse branches of different scales and complexities to enrich the feature space, including sequences of convolutions, multi-scale convolutions, and average pooling. After training, a DBB can be equivalently converted into a single conv layer for deployment. Unlike the advancements of novel ConvNet architectures, DBB complicates the training-time microstructure while maintaining the macro architecture, so that it can be used as a drop-in replacement for regular conv layers of any architecture. In this way, the model can be trained to reach a higher level of performance and then transformed into the original inference-time structure for inference. DBB improves ConvNets on image classification (up to 1.9% higher top-1 accuracy on ImageNet), object detection and semantic segmentation. The PyTorch code and models are released at https://github.com/DingXiaoH/DiverseBranchBlock.
****
Post-Hoc Uncertainty Calibration for Domain Drift Scenarios
Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, Florian Buettner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10124-10132
We address the problem of uncertainty calibration. While standard deep neural networks typically yield uncalibrated predictions, calibrated confidence scores that are representative of the true likelihood of a prediction can be achieved using post-hoc calibration methods. However, to date, the focus of these approaches has been on in-domain calibration. Our contribution is two-fold. First, we show that existing post-hoc calibration methods yield highly over-confident predictions under domain shift. Second, we introduce a simple strategy where perturbations are applied to samples in the validation set before performing the post-hoc calibration step. In extensive experiments, we demonstrate that this perturbation step results in substantially better calibration under domain shift on a wide range of architectures and modelling tasks.
****
Slimmable Compressive Autoencoders for Practical Neural Image Compression
Fei Yang, Luis Herranz, Yongmei Cheng, Mikhail G. Mozerov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4998-5007
Neural image compression leverages deep neural networks to outperform traditional image codecs in rate-distortion performance. However, the resulting models are also heavy, computationally demanding and generally optimized for a single rate, limiting their practical use. Focusing on practical image compression, we propose slimmable compressive autoencoders (SlimCAEs), where rate (R) and distortion (D) are jointly optimized for different capacities. Once trained, encoders and decoders can be executed at different capacities, leading to different rates and complexities. We show that a successful implementation of SlimCAEs requires suitable capacity-specific RD tradeoffs. Our experiments show that SlimCAEs are hig

hly flexible models that provide excellent rate-distortion performance, variable rate, and dynamic adjustment of memory, computational cost and latency, thus addressing the main requirements of practical image compression.
****
Function4D: Real-Time Human Volumetric Capture From Very Sparse Consumer RGBD Sensors
Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5746-5756
Human volumetric capture is a long-standing topic in computer vision and computer graphics. Although high-quality results can be achieved using sophisticated off-line systems, real-time human volumetric capture of complex scenarios, especially using light-weight setups, remains challenging. In this paper, we propose a human volumetric capture method that combines temporal volumetric fusion and deep implicit functions. To achieve high-quality and temporal-continuous reconstruction, we propose dynamic sliding fusion to fuse neighboring depth observations together with topology consistency. Moreover, for detailed and complete surface generation, we propose detail-preserving deep implicit functions for RGBD input which can not only preserve the geometric details on the depth inputs but also generate more plausible texturing results. Results and experiments show that our method outperforms existing methods in terms of view sparsity, generalization capacity, reconstruction quality, and run-time efficiency.
****
LAU-Net: Latitude Adaptive Upscaling Network for Omnidirectional Image Super-Resolution
Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, Li Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9189-9198
The omnidirectional images (ODIs) are usually at low-resolution, due to the constraints of collection, storage and transmission. The traditional two-dimensional (2D) image super-resolution methods are not effective for spherical ODIs, because ODIs tend to have non-uniformly distributed pixel density and varying texture complexity across latitudes. In this work, we propose a novel latitude adaptive upscaling network (LAU-Net) for ODI super-resolution, which allows pixels at different latitudes to adopt distinct upscaling factors. Specifically, we introduce a Laplacian multi-level separation architecture to split an ODI into different latitude bands, and hierarchically upscale them with different factors. In addition, we propose a deep reinforcement learning scheme with a latitude adaptive reward, in order to automatically select optimal upscaling factors for different latitude bands. To the best of our knowledge, LAU-Net is the first attempt to consider the latitude difference for ODI super-resolution. Extensive results demonstrate that our LAU-Net significantly advances the super-resolution performance for ODIs.
****
UP-DETR: Unsupervised Pre-Training for Object Detection With Transformers
Zhigang Dai, Bolun Cai, Yugeng Lin, Junying Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1601-1610
Object detection with transformers (DETR) reaches competitive performance with Faster R-CNN via a transformer encoder-decoder architecture. Inspired by the great success of pre-training transformers in natural language processing, we propose a pretext task named random query patch detection to Unsupervisedly Pre-train DETR (UP-DETR) for object detection. Specifically, we randomly crop patches from the given image and then feed them as queries to the decoder. The model is pre-trained to detect these query patches from the original image. During the pre-training, we address two critical issues: multi-task learning and multi-query localization. (1) To trade off classification and localization preferences in the pretext task, we freeze the CNN backbone and propose a patch feature reconstruction branch which is jointly optimized with patch detection. (2) To perform multi-query localization, we introduce UP-DETR from single-query patch and extend it to multi-query patches with object query shuffle and attention mask. In our experi

ments, UP-DETR significantly boosts the performance of DETR with faster convergence and higher average precision on object detection, one-shot detection and panoptic segmentation. Code and pre-training models: https://github.com/dddzg/up-detr.
****
Self-Attention Based Text Knowledge Mining for Text Detection
Qi Wan, Haoqin Ji, Linlin Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5983-5992
Pre-trained models play an important role in deep learning based text detectors. However, most methods ignore the gap between natural images and scene text images and directly apply ImageNet for pre-training. To address such a problem, some of them firstly pre-train the model using a large amount of synthetic data and then fine-tune it on target datasets, which is task-specific and has limited generalization capability. In this paper, we focus on providing general pre-trained models for text detectors. Considering the importance of exploring text contents for text detection, we propose STKM (Self-attention based Text Knowledge Mining), which consists of a CNN Encoder and a Self-attention Decoder, to learn general prior knowledge for text detection from SynthText. Given only image level text labels, Self-attention Decoder directly decodes features extracted from CNN Encoder to texts without requirement of detection, which guides the CNN backbone to explicitly learn discriminative semantic representations ignored by previous approaches. After that, the text knowledge learned by the backbone can be transferred to various text detectors to significantly improve their detection performance (e.g., 5.89% higher F-measure for EAST on ICDAR15 dataset) without bells and whistles. Pre-trained model is available at: https://github.com/CVI-SZU/STKM
****
Image De-Raining via Continual Learning
Man Zhou, Jie Xiao, Yifan Chang, Xueyang Fu, Aiping Liu, Jinshan Pan, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4907-4916
While deep convolutional neural networks (CNNs) have achieved great success on image de-raining task, most existing methods can only learn fixed mapping rules between paired rainy/clean images on a single dataset. This limits their applications in practical situations with multiple and incremental datasets where the mapping rules may change for different types of rain streaks. However, the catastrophic forgetting of traditional deep CNN model challenges the design of generalized framework for multiple and incremental datasets. A strategy of sharing the network structure but independently updating and storing the network parameters on each dataset has been developed as a potential solution. Nevertheless, this strategy is not applicable to compact systems as it dramatically increases the overall training time and parameter space. To alleviate such limitation, in this study, we propose a parameter importance guided weights modification approach, named PIGWM. Specifically, with new dataset (e.g. new rain dataset), the well-trained network weights are updated according to their importance evaluated on previous training dataset. With extensive experimental validation, we demonstrate that a single network with a single parameter set of our proposed method can process multiple rain datasets almost without performance degradation. The proposed model is capable of achieving superior performance on both inhomogeneous and incremental datasets, and is promising for highly compact systems to gradually learn myriad regularities of the different types of rain streaks. The results indicate that our proposed method has great potential for other computer vision tasks with dynamic learning environments.
****
Layer-Wise Searching for 1-Bit Detectors
Sheng Xu, Junhe Zhao, Jinhu Lu, Baochang Zhang, Shumin Han, David Doermann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5682-5691
1-bit detectors show great promise for resource-constrained embedded devices but often suffer from a significant performance gap compared with their real-valued counterparts. The primary reason lies in the layer-wise error during binarizati

on. This paper presents a layer-wise search (LWS) strategy to generate 1-bit detectors that maintain a performance very close to the original real-valued model. The approach introduces angular and amplitude angular error loss functions to increase detector capacity. At each layer, it exploits a differentiable binarization search (DBS) to minimize the angular error in a student-teacher framework. It then fine-tunes the scale parameter of that layer to reduce the amplitude error. Extensive experiments show that LWS-Det outperforms state-of-the-art 1-bit detectors by a considerable margin on the PASCAL VOC and COCO datasets. For example, the LWS-Det achieves 1-bit Faster-RCNN with ResNet-34 backbone within 2.0% mAP of its real-valued counterpart on the PASCAL VOC dataset.
****
Distilling Audio-Visual Knowledge by Compositional Contrastive Learning
Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7016-7025
Having access to multi-modal cues (e.g. vision and audio) empowers some cognitive tasks to be done faster compared to learning from a single modality. In this work, we propose to transfer knowledge across heterogeneous modalities, even though these data modalities may not be semantically correlated. Rather than directly aligning the representations of different modalities, we compose audio, image, and video representations across modalities to uncover the richer multi-modal knowledge. Our main idea is to learn a compositional embedding that closes the cross-modal semantic gap and captures the task-relevant semantics, which facilitates pulling together representations across modalities by compositional contrastive learning. We establish a new, comprehensive multi-modal distillation benchmark on three video datasets: UCF101, ActivityNet, and VGGSound. Moreover, we demonstrate that our model significantly outperforms a variety of existing knowledge distillation methods in transferring audio-visual knowledge to improve video representation learning.
****
Unsupervised Visual Attention and Invariance for Reinforcement Learning
Xudong Wang, Long Lian, Stella X. Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6677-6687
The vision-based reinforcement learning (RL) has achieved tremendous success. However, generalizing vision-based RL policy to unknown test environments still remains as a challenging problem. Unlike previous works that focus on training a universal RL policy that is invariant to discrepancies between test and training environment, we focus on developing an independent module to disperse interference factors irrelevant to the task, thereby providing ""clean"" observations for the RL policy. The proposed unsupervised visual attention and invariance method (VAI) contains three key components: 1) an unsupervised keypoint detection model which captures semantically meaningful keypoints in observations; 2) an unsupervised visual attention module which automatically generates the distraction-invariant attention mask for each observation; 3) a self-supervised adapter for visual distraction invariance which reconstructs distraction-invariant attention mask from observations with artificial disturbances generated by a series of foreground and background augmentations. All components are optimized in an unsupervised way, without manual annotation or access to environment internals, and only the adapter is used during inference time to provide distraction-free observations to RL policy. VAI empirically shows powerful generalization capabilities and significantly outperforms current state-of-the-art (SOTA) method by 15% 49% in DeepMind Control suite benchmark and 61% 229% in our proposed robot manipulation benchmark, in term of cumulative rewards per episode.
****
CRFace: Confidence Ranker for Model-Agnostic Face Detection Refinement
Noranart Vesdapunt, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1674-1684
Face detection is a fundamental problem for many downstream face applications, and there is a rising demand for faster, more accurate yet support for higher resolution face detectors. Recent smartphones can record a video in 8K resolution,

but many of the existing face detectors still fail due to the anchor size and training data. We analyze the failure cases and observe a large number of correct predicted boxes with incorrect confidences. To calibrate these confidences, we propose a confidence ranking network with a pairwise ranking loss to re-rank the predicted confidences locally within the same image. Our confidence ranker is model-agnostic, so we can augment the data by choosing the pairs from multiple face detectors during the training, and generalize to a wide range of face detectors during the testing. On WiderFace, we achieve the highest AP on the single-scale, and our AP is competitive with the previous multi-scale methods while being significantly faster. On 8K resolution, our method solves the GPU memory issue and allows us to indirectly train on 8K. We collect 8K resolution test set to show the improvement, and we will release our test set as a new benchmark for future research.
****
Semantic Audio-Visual Navigation
Changan Chen, Ziad Al-Halah, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15516-15525
Recent work on audio-visual navigation assumes a constantly-sounding target and restricts the role of audio to signaling the target's position. We introduce semantic audio-visual navigation, where objects in the environment make sounds consistent with their semantic meaning (e.g., toilet flushing, door creaking) and acoustic events are sporadic or short in duration. We propose a transformer-based model to tackle this new semantic AudioGoal task, incorporating an inferred goal descriptor that captures both spatial and semantic properties of the target. Our model's persistent multimodal memory enables it to reach the goal even long after the acoustic event stops. In support of the new task, we also expand the SoundSpaces audio simulations to provide semantically grounded sounds for an array of objects in Matterport3D. Our method strongly outperforms existing audio-visual navigation methods by learning to associate semantic, acoustic, and visual cues. Project page: http://vision.cs.utexas.edu/projects/semantic-audio-visual-navigation.
****
Humble Teachers Teach Better Students for Semi-Supervised Object Detection
Yihe Tang, Weifeng Chen, Yijun Luo, Yuting Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3132-3141
We propose a semi-supervised approach for contemporary object detectors following the teacher-student dual model framework. Our method is featured with 1) the exponential moving averaging strategy to update the teacher from the student online, 2) using plenty of region proposals and soft pseudo-labels as the student's training targets, and 3) a light-weighted detection-specific data ensemble for the teacher to generate more reliable pseudo labels. Compared to the recent state-of-the-art - STAC, which uses hard labels on sparsely selected hard pseudo samples, the teacher in our model exposes richer information to the student with soft-labels on many proposals. Our model achieves COCO-style AP of 53.04% on VOC07 val set, 8.4% better than STAC, when using VOC12 as unlabeled data. On MS-COCO, it outperforms prior work when only a small percentage of data is taken as labeled. It also reaches 53.8% AP on MS-COCO test-dev with 3.1% gain over the fully supervised ResNet-152 cascaded R-CNN, by tapping into unlabeled data of a similar size to the labeled data.
****
One Shot Face Swapping on Megapixels
Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, Zhenan Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4834-4844
Face swapping has both positive applications such as entertainment, human-computer interaction, etc., and negative applications such as DeepFake threats to politics, economics, etc. Nevertheless, it is necessary to understand the scheme of advanced methods for high-quality face swapping and generate enough and representative face swapping images to train DeepFake detection algorithms. This paper proposes the first Megapixel level method for one shot Face Swapping (or MegaFS f

or short). Firstly, MegaFS organizes face representation hierarchically by the proposed Hierarchical Representation Face Encoder (HieRFE) in an extended latent space to maintain more facial details, rather than compressed representation in previous face swapping methods. Secondly, a carefully designed Face Transfer Module (FTM) is proposed to transfer the identity from a source image to the target by a non-linear trajectory without explicit feature disentanglement. Finally, the swapped faces can be synthesized by StyleGAN2 with the benefits of its training stability and powerful generative capability. Each part of MegaFS can be trained separately so the requirement of our model for GPU memory can be satisfied for megapixel face swapping. In summary, complete face representation, stable training, and limited memory usage are the three novel contributions to the success of our method. Extensive experiments demonstrate the superiority of MegaFS and the first megapixel level face swapping database is released for research on DeepFake detection and face image editing in the public domain.
****
CDFI: Compression-Driven Network Design for Frame Interpolation
Tianyu Ding, Luming Liang, Zhihui Zhu, Ilya Zharkov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8001-8011
DNN-based frame interpolation--that generates the intermediate frames given two consecutive frames--typically relies on heavy model architectures with a huge number of features, preventing them from being deployed on systems with limited resources, e.g., mobile devices. We propose a compression-driven network design for frame interpolation (CDFI), that leverages model pruning through sparsity-inducing optimization to significantly reduce the model size while achieving superior performance. Concretely, we first compress the recently proposed AdaCoF model and show that a 10X compressed AdaCoF performs similarly as its original counterpart; then we further improve this compressed model by introducing a multi-resolution warping module, which boosts visual consistencies with multi-level details. As a consequence, we achieve a significant performance gain with only a quarter in size compared with the original AdaCoF. Moreover, our model performs favorably against other state-of-the-arts in a broad range of datasets. Finally, the proposed compression-driven framework is generic and can be easily transferred to other DNN-based frame interpolation algorithm. Our source code is available at https://github.com/tding1/CDFI.
****
PAConv: Position Adaptive Convolution With Dynamic Kernel Assembling on Point Clouds
Mutian Xu, Runyu Ding, Hengshuang Zhao, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3173-3182
We introduce Position Adaptive Convolution (PAConv), a generic convolution operation for 3D point cloud processing. The key of PAConv is to construct the convolution kernel by dynamically assembling basic weight matrices stored in Weight Bank, where the coefficients of these weight matrices are self-adaptively learned from point positions through ScoreNet. In this way, the kernel is built in a data-driven manner, endowing PAConv with more flexibility than 2D convolutions to better handle the irregular and unordered point cloud data. Besides, the complexity of the learning process is reduced by combining weight matrices instead of brutally predicting kernels from point positions. Furthermore, different from the existing point convolution operators whose network architectures are often heavily engineered, we integrate our PAConv into classical MLP-based point cloud pipelines without changing network configurations. Even built on simple networks, our method still approaches or even surpasses the state-of-the-art models, and significantly improves baseline performance on both classification and segmentation tasks, yet with decent efficiency. Thorough ablation studies and visualizations are provided to understand PAConv. Code is released on https://github.com/CVMI-Lab/PAConv.
****
End-to-End Object Detection With Fully Convolutional Network

Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15849-15858
Mainstream object detectors based on the fully convolutional network has achieved impressive performance. While most of them still need a hand-designed non-maximum suppression (NMS) post-processing, which impedes fully end-to-end training. In this paper, we give the analysis of discarding NMS, where the results reveal that a proper label assignment plays a crucial role. To this end, for fully convolutional detectors, we introduce a Prediction-aware One-To-One (POTO) label assignment for classification to enable end-to-end detection, which obtains comparable performance with NMS. Besides, a simple 3D Max Filtering (3DMF) is proposed to utilize the multi-scale features and improve the discriminability of convolutions in the local region. With these techniques, our end-to-end framework achieves competitive performance against many state-of-the-art detectors with NMS on COCO and CrowdHuman datasets. The code is available at https://github.com/Megvii-BaseDetection/DeFCN.
****
Efficient Initial Pose-Graph Generation for Global SfM
Daniel Barath, Dmytro Mishkin, Ivan Eichhardt, Ilia Shipachev, Jiri Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14546-14555
We propose ways to speed up the initial pose-graph generation for global Structure-from-Motion algorithms. To avoid forming tentative point correspondences by FLANN and geometric verification by RANSAC, which are the most time-consuming steps of the pose-graph creation, we propose two new methods -- built on the fact that image pairs usually are matched consecutively. Thus, candidate relative poses can be recovered from paths in the partly-built pose-graph. We propose a heuristic for the A* traversal, considering global similarity of images and the quality of the pose-graph edges. Given a relative pose from a path, descriptor-based feature matching is made "light-weight" by exploiting the known epipolar geometry. To speed up PROSAC-based sampling when RANSAC is applied, we propose a third method to order the correspondences by their inlier probabilities from previous estimations. The algorithms are tested on 402130 image pairs from the 1DSfM dataset and they speed up the feature matching 17 times and pose estimation 5 times. The source code will be made public.
****
Representative Batch Normalization With Feature Calibration
Shang-Hua Gao, Qi Han, Duo Li, Ming-Ming Cheng, Pai Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8669-8679
Batch Normalization (BatchNorm) has become the default component in modern neural networks to stabilize training. In BatchNorm, centering and scaling operations, along with mean and variance statistics, are utilized for feature standardization over the batch dimension. The batch dependency of BatchNorm enables stable training and better representation of the network, while inevitably ignores the representation differences among instances. We propose to add a simple yet effective feature calibration scheme into the centering and scaling operations of BatchNorm, enhancing the instance-specific representations with the negligible computational cost. The centering calibration strengthens informative features and reduces noisy features. The scaling calibration restricts the feature intensity to form a more stable feature distribution. Our proposed variant of BatchNorm, namely Representative BatchNorm, can be plugged into existing methods to boost the performance of various tasks such as classification, detection, and segmentation. The source code is available in http://mmcheng.net/rbn.
****
VarifocalNet: An IoU-Aware Dense Object Detector
Haoyang Zhang, Ying Wang, Feras Dayoub, Niko Sunderhauf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8514-8523
Accurately ranking the vast number of candidate detections is crucial for dense

object detectors to achieve high performance. Prior work uses the classification score or a combination of classification and predicted localization scores to rank candidates. However, neither option results in a reliable ranking, thus degrading detection performance. In this paper, we propose to learn an Iou-Aware Classification Score (IACS) as a joint representation of object presence confidence and localization accuracy. We show that dense object detectors can achieve a more accurate ranking of candidate detections based on the IACS. We design a new loss function, named Varifocal Loss, to train a dense object detector to predict the IACS, and propose a new star-shaped bounding box feature representation for IACS prediction and bounding box refinement. Combining these two new components and a bounding box refinement branch, we build an IoU-aware dense object detector based on the FCOS+ATSS architecture, that we call VarifocalNet or VFNet for short. Extensive experiments on MS COCO show that our VFNet consistently surpasses the strong baseline by 2.0 AP with different backbones. Our best model VFNet-X-1200 with Res2Net-101-DCN achieves a single-model single-scale AP of 55.1 on COCO test-dev, which is state-of-the-art among various object detectors. Code is available at: https://github.com/hyz-xmaster/VarifocalNet.
****
Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation
Youngmin Oh, Beomjun Kim, Bumsub Ham; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6913-6922
We address the problem of weakly-supervised semantic segmentation (WSSS) using bounding box annotations. Although object bounding boxes are good indicators to segment corresponding objects, they do not specify object boundaries, making it hard to train convolutional neural networks (CNNs) for semantic segmentation. We find that background regions are perceptually consistent in part within an image, and this can be leveraged to discriminate foreground and background regions inside object bounding boxes. To implement this idea, we propose a novel pooling method, dubbed background-aware pooling (BAP), that focuses more on aggregating foreground features inside the bounding boxes using attention maps. This allows to extract high-quality pseudo segmentation labels to train CNNs for semantic segmentation, but the labels still contain noise especially at object boundaries. To address this problem, we also introduce a noise-aware loss (NAL) that makes the networks less susceptible to incorrect labels. Experimental results demonstrate that learning with our pseudo labels already outperforms state-of-the-art weakly- and semi-supervised methods on the PASCAL VOC 2012 dataset, and the NAL further boosts the performance.
****
Abstract Spatial-Temporal Reasoning via Probabilistic Abduction and Execution
Chi Zhang, Baoxiong Jia, Song-Chun Zhu, Yixin Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9736-9746
Spatial-temporal reasoning is a challenging task in Artificial Intelligence (AI) due to its demanding but unique nature: a theoretic requirement on representing and reasoning based on spatial-temporal knowledge in mind, and an applied requirement on a high-level cognitive system capable of navigating and acting in space and time. Recent works have focused on an abstract reasoning task of this kind---Raven's Progressive Matrices (RPM). Despite the encouraging progress on RPM that achieves human-level performance in terms of accuracy, modern approaches have neither a treatment of human-like reasoning on generalization, nor a potential to generate answers. To fill in this gap, we propose a neuro-symbolic Probabilistic Abduction and Execution (PrAE) learner; central to the PrAE learner is the process of probabilistic abduction and execution on a probabilistic scene representation, akin to the mental manipulation of objects. Specifically, we disentangle perception and reasoning from a monolithic model. The neural visual perception frontend predicts objects' attributes, later aggregated by a scene inference engine to produce a probabilistic scene representation. In the symbolic logical reasoning backend, the PrAE learner uses the representation to abduce the hidden rules. An answer is predicted by executing the rules on the probabilistic representation. The entire system is trained end-to-end in an analysis-by-synthesis ma

nner without any visual attribute annotations. Extensive experiments demonstrate that the PrAE learner improves cross-configuration generalization and is capable of rendering an answer, in contrast to prior works that merely make a categorical choice from candidates.
****
Reducing Domain Gap by Reducing Style Bias
Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, Donggeun Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8690-8699
Convolutional Neural Networks (CNNs) often fail to maintain their performance when they confront new test domains, which is known as the problem of domain shift. Recent studies suggest that one of the main causes of this problem is CNNs' strong inductive bias towards image styles (i.e. textures) which are sensitive to domain changes, rather than contents (i.e. shapes). Inspired by this, we propose to reduce the intrinsic style bias of CNNs to close the gap between domains. Our Style-Agnostic Networks (SagNets) disentangle style encodings from class categories to prevent style biased predictions and focus more on the contents. Extensive experiments show that our method effectively reduces the style bias and makes the model more robust under domain shift. It achieves remarkable performance improvements in a wide range of cross-domain tasks including domain generalization, unsupervised domain adaptation, and semi-supervised domain adaptation on multiple datasets.
****
Efficient Regional Memory Network for Video Object Segmentation
Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, Wenxiu Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1286-1295
Recently, several Space-Time Memory based networks have shown that the object cues (e.g. video frames as well as the segmented object masks) from the past frames are useful for segmenting objects in the current frame. However, these methods exploit the information from the memory by global-to-global matching between the current and past frames, which lead to mismatching to similar objects and high computational complexity. To address these problems, we propose a novel local-to-local matching solution for semi-supervised VOS, namely Regional Memory Network (RMNet). In RMNet, the precise regional memory is constructed by memorizing local regions where the target objects appear in the past frames. For the current query frame, the query regions are tracked and predicted based on the optical flow estimated from the previous frame. The proposed local-to-local matching effectively eliminates the ambiguity of similar objects in both memory and query frames, which allows the information to be passed from the regional memory to the query region efficiently and effectively. Experimental results indicate that the proposed RMNet performs favorably against state-of-the-art methods on the DAVIS and YouTube-VOS datasets.
****
Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-Localization in Large Scenes From Body-Mounted Sensors
Vladimir Guzov, Aymen Mir, Torsten Sattler, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4318-4329
We introduce (HPS) Human POSEitioning System, a method to recover the full 3D pose of a human registered with a 3D scan of the surrounding environment using wearable sensors. Using IMUs attached at the body limbs and a head mounted camera looking outwards, HPS fuses camera based self-localization with IMU-based human body tracking. The former provides drift-free but noisy position and orientation estimates while the latter is accurate in the short-term but subject to drift over longer periods of time. We show that our optimization-based integration exploits the benefits of the two, resulting in pose accuracy free of drift. Furthermore, we integrate 3D scene constraints into our optimization, such as foot contact with the ground, resulting in physically plausible motion. HPS complements more common third-person-based 3D pose estimation methods. It allows capturing larg

er recording volumes and longer periods of motion, and could be used for VR/AR applications where humans interact with the scene without requiring direct line of sight with an external camera, or to train agents that navigate and interact with the environment based on first-person visual input, like real humans. With HPS, we recorded a dataset of humans interacting with large 3D scenes (300-1000 sq.m) consisting of 7 subjects and more than 3 hours of diverse motion. The dataset, code and video will be available on the project page: http://virtualhumans.mpi-inf.mpg.de/hps/.
****
Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection
Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, Marios Savvides; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8782-8791
Few-shot object detection is an imperative and long-lasting problem due to the inherent long-tail distribution of real-world data. Its performance is largely affected by the data scarcity of novel classes. But the semantic relation between the novel classes and the base classes is constant regardless of the data availability. In this work, we investigate utilizing this semantic relation together with the visual information and introduce explicit relation reasoning into the learning of novel object detection. Specifically, we represent each class concept by a semantic embedding learned from a large corpus of text. The detector is trained to project the image representations of objects into this embedding space. We also identify the problems of trivially using the raw embeddings with a heuristic knowledge graph and propose to augment the embeddings with a dynamic relation graph. As a result, our few-shot detector, termed SRR-FSD, is robust and stable to the variation of shots of novel objects. Experiments show that SRR-FSD can achieve competitive results at higher shots, and more importantly, a significantly better performance given both lower explicit and implicit shots. The benchmark protocol with implicit shots removed from the pretrained classification dataset can serve as a more realistic setting for future research.
****
Online Multiple Object Tracking With Cross-Task Synergy
Song Guo, Jingya Wang, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8136-8145
Modern online multiple object tracking (MOT) methods usually focus on two directions to improve tracking performance. One is to predict new positions in an incoming frame based on tracking information from previous frames, and the other is to enhance data association by generating more discriminative identity embeddings. Some works combined both directions within one framework but handled them as two individual tasks, thus gaining little mutual benefits. In this paper, we propose a novel unified model with synergy between position prediction and embedding association. The two tasks are linked by temporal-aware target attention and distractor attention, as well as identity-aware memory aggregation model. Specifically, the attention modules can make the prediction focus more on targets and less on distractors, therefore more reliable embeddings can be extracted accordingly for association. On the other hand, such reliable embeddings can boost identity-awareness through memory aggregation, hence strengthen attention modules and suppress drifts. In this way, the synergy between position prediction and embedding association is achieved, which leads to strong robustness to occlusions. Extensive experiments demonstrate the superiority of our proposed model over a wide range of existing methods on MOTChallenge benchmarks. Our code and models are publicly available at https://github.com/songguocode/TADAM
****
Discovering Relationships Between Object Categories via Universal Canonical Maps
Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 404-413
We tackle the problem of learning the geometry of multiple categories of deformable objects jointly. Recent work has shown that it is possible to learn a unified dense pose predictor for several categories of related objects. However, train

ing such models requires to initialize inter-category correspondences by hand. This is suboptimal and the resulting models fail to maintain correct correspondences as individual categories are learned. In this paper, we show that improved correspondences can be learned automatically as a natural byproduct of learning category-specific dense pose predictors. To do this, we express correspondences between different categories and between images and categories using a unified embedding. Then, we use the latter to enforce two constraints: symmetric inter-category cycle consistency and a new asymmetric image-to-category cycle consistency. Without any manual annotations for the inter-category correspondences, we obtain state-of-the-art alignment results, outperforming dedicated methods for matching 3D shapes. Moreover, the new model is also better at the task of dense pose prediction than prior work.
****
Prior Based Human Completion
Zibo Zhao, Wen Liu, Yanyu Xu, Xianing Chen, Weixin Luo, Lei Jin, Bohui Zhu, Tong Liu, Binqiang Zhao, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7951-7961
We study a very challenging task, human image completion, which tries to recover the human body part with a reasonable human shape from the corrupted region. Since each human body part is unique, it is infeasible to restore the missing part by borrowing textures from other visible regions. Thus, we propose two types of learned priors to compensate for the damaged region. One is a structure prior, it uses a human parsing map to represent the human body structure. The other is a structure-texture correlation prior. It learns a structure and a texture memory bank, which encodes the common body structures and texture patterns, respectively. With the aid of these memory banks, the model could utilize the visible pattern to query and fetch a similar structure and texture pattern to introduce additional reasonable structures and textures for the corrupted region. Besides, since multiple potential human shapes are underlying the corrupted region, we propose multi-scale structure discriminators to further restore a plausible topological structure. Experiments on various large-scale benchmarks demonstrate the effectiveness of our proposed method.
****
Neural Response Interpretation Through the Lens of Critical Pathways
Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Christian Rupprecht, Seong Tae Kim, Nassir Navab; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13528-13538
Is critical input information encoded in specific sparse pathways within the neural network? In this work, we discuss the problem of identifying these critical pathways and subsequently leverage them for interpreting the network's response to an input. The pruning objective --- selecting the smallest group of neurons for which the response remains equivalent to the original network --- has been previously proposed for identifying critical pathways. We demonstrate that sparse pathways derived from pruning do not necessarily encode critical input information. To ensure sparse pathways include critical fragments of the encoded input information, we propose pathway selection via neurons' contribution to the response. We proceed to explain how critical pathways can reveal critical input features. We prove that pathways selected via neuron contribution are locally linear (in an L2-ball), a property that we use for proposing a feature attribution method: "pathway gradient". We validate our interpretation method using mainstream evaluation experiments. The validation of pathway gradient interpretation method further confirms that selected pathways using neuron contributions correspond to critical input features. The code is publicly available.
****
Rethinking and Improving the Robustness of Image Style Transfer
Pei Wang, Yijun Li, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 124-133
Extensive research in neural style transfer methods has shown that the correlation between features extracted by a pre-trained VGG network has remarkable ability to capture the visual style of an image. Surprisingly, however, this stylizati

on quality is not robust and often degrades significantly when applied to featur
es from more advanced and lightweight networks, such as those in the ResNet fami
ly. By performing extensive experiments with different network architectures, we
find that residual connections, which represent the main architectural differen
ce between VGG and ResNet, produce feature maps of small entropy, which are not
suitable for style transfer. To improve the robustness of the ResNet architectur
e, we then propose a simple yet effective solution based on a softmax transforma
tion of the feature activations that enhances their entropy. Experimental result
s demonstrate that this small magic can greatly improve the quality of stylizati
on results, even for networks with random weights. This suggests that the archit
ecture used for feature extraction is more important than the use of learned wei
ghts for the task of style transfer.
****
FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding
Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, Chi Zhang; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7352-
7362
Emerging interests have been brought to recognize previously unseen objects give
n very few training examples, known as few-shot object detection (FSOD). Recent
researches demonstrate that good feature embedding is the key to reach favorable
few-shot learning performance. We observe object proposals with different Inter
section-of-Union (IoU) scores are analogous to the intra-image augmentation used
in contrastive visual representation learning. And we exploit this analogy and
incorporate supervised contrastive learning to achieve more robust objects repre
sentations in FSOD. We present Few-Shot object detection via Contrastive proposa
ls Encoding (FSCE), a simple yet effective approach to learning contrastive-awar
e object proposal encodings that facilitate the classification of detected objec
ts. We notice the degradation of average precision (AP) for rare objects mainly
comes from misclassifying novel instances as confusable classes. And we ease the
misclassification issues by promoting instance level intra-class compactness an
d inter-class variance via our contrastive proposal encoding loss (CPE loss). Ou
r design outperforms current state-of-the-art works in any shot and all data spl
its, with up to +8.8% on standard benchmark PASCAL VOC and +2.7% on challenging
COCO benchmark. Code is available at: https://github.com/ MegviiDetection/FSCE.
****
Cross-Domain Similarity Learning for Face Recognition in Unseen Domains
Masoud Faraki, Xiang Yu, Yi-Hsuan Tsai, Yumin Suh, Manmohan Chandraker; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2021, pp. 15292-15301
Face recognition models trained under the assumption of identical training and t
est distributions often suffer from poor generalization when faced with unknown
variations, such as a novel ethnicity or unpredictable individual make-ups durin
g test time. In this paper, we introduce a novel cross-domain metric learning lo
ss, which we dub Cross-Domain Triplet (CDT) loss, to improve face recognition in
unseen domains. The CDT loss encourages learning semantically meaningful featur
es by enforcing compact feature clusters of identities from one domain, where th
e compactness is measured by underlying similarity metrics that belong to anothe
r training domain with different statistics. Intuitively, it discriminatively co
rrelates explicit metrics derived from one domain, with triplet samples from ano
ther domain in a unified loss function to be minimized within a network, which l
eads to better alignment of the training domains. The network parameters are fur
ther enforced to learn generalized features under domain shift, in a model-agnos
tic learning pipeline. Unlike the recent work of Meta Face Recognition, our meth
od does not require careful hard-pair sample mining and filtering strategy durin
g training. Extensive experiments on various face recognition benchmarks show th
e superiority of our method in handling variations, compared to baseline methods
and the state-of-the-arts.
****
Learning 3D Shape Feature for Texture-Insensitive Person Re-Identification
Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, Wei-S

hi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8146-8155
It is well acknowledged that person re-identification (person ReID) highly relies on visual texture information like clothing. Despite significant progress has been made in recent years, texture-confusing situations like clothing changing and persons wearing the same clothes receive little attention from most existing ReID methods. In this paper, rather than relying on texture based information, we propose to improve the robustness of person ReID against clothing texture by exploiting the information of a person's 3D shape. Existing shape learning schemas for person ReID either ignore the 3D information of a person, or require extra physical devices to collect 3D source data. Differently, we propose a novel ReID learning framework that directly extracts a texture-insensitive 3D shape embedding from a 2D image by adding 3D body reconstruction as an auxiliary task and regularization, called 3D Shape Learning (3DSL). The 3D reconstruction based regularization forces the ReID model to decouple the 3D shape information from the visual texture, and acquire discriminative 3D shape ReID features. To solve the problem of lacking 3D ground truth, we design an adversarial self-supervised projection (ASSP) model, performing 3D reconstruction without ground truth. Extensive experiments on common ReID datasets and texture-confusing datasets validate the effectiveness of our model.
****
Virtual Fully-Connected Layer: Training a Large-Scale Face Recognition Dataset With Limited Computational Resources
Pengyu Li, Biao Wang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13315-13324
Recently, deep face recognition has achieved significant progress because of Convolutional Neural Networks (CNNs) and large-scale datasets. However, training CNNs on a large-scale face recognition dataset with limited computational resources is still a challenge. This is because the classification paradigm needs to train a fully connected layer as the category classifier, and its parameters will be in the hundreds of millions if the training dataset contains millions of identities. This requires many computational resources, such as GPU memory. The metric learning paradigm is an economical computation method, but its performance is greatly inferior to that of the classification paradigm. To address this challenge, we propose a simple but effective CNN layer called the Virtual fully connected (Virtual FC) layer to reduce the computational consumption of the classification paradigm. Without bells and whistles, the proposed Virtual FC reduces the parameters by more than 100 times with respect to the fully connected layer and achieves competitive performance on mainstream face recognition evaluation datasets. Moreover, the performance of our Virtual FC layer on the evaluation datasets is superior to that of the metric learning paradigm by a significant margin. Our code will be released in hopes of disseminating our idea to other domains.
****
Multi-Person Implicit Reconstruction From a Single Image
Armin Mustafa, Akin Caliskan, Lourdes Agapito, Adrian Hilton; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14474-14483
We present a new end-to-end learning framework to obtain detailed and spatially coherent reconstructions of multiple people from a single image. Existing multi-person methods suffer from two main drawbacks: they are often model-based and therefore cannot capture accurate 3D models of people with loose clothing and hair; or they require manual intervention to resolve occlusions or interactions. Our method addresses both limitations by introducing the first end-to-end learning approach to perform model-free implicit reconstruction for realistic 3D capture of multiple clothed people in arbitrary poses (with occlusions) from a single image. Our network simultaneously estimates the 3D geometry of each person and their 6DOF spatial locations, to obtain a coherent multi-human reconstruction. In addition, we introduce a new synthetic dataset that depicts images with a varying number of inter-occluded humans in a variety of clothing and hair. We demonstrate robust, high-resolution reconstructions on images of multiple humans with com

plex occlusions, loose clothing and a large variety of poses, and scenes. Our quantitative evaluation on both synthetic and real world datasets demonstrates state-of-the-art performance with significant improvements in the accuracy and completeness of the reconstructions over competing approaches.
****
OPANAS: One-Shot Path Aggregation Network Architecture Search for Object Detection
Tingting Liang, Yongtao Wang, Zhi Tang, Guosheng Hu, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10195-10203
Recently, neural architecture search (NAS) has been exploited to design feature pyramid networks (FPNs) and achieved promising results for visual object detection. Encouraged by the success, we propose a novel One-Shot Path Aggregation Network Architecture Search (OPANAS) algorithm, which significantly improves both searching efficiency and detection accuracy. Specifically, we first introduce six heterogeneous information paths to build our search space, namely top-down, bottom-up, fusing-splitting, scale-equalizing, skip-connect, and none. Second, we propose a novel search space of FPNs, in which each FPN candidate is represented by a densely-connected directed acyclic graph (each node is a feature pyramid and each edge is one of the six heterogeneous information paths). Third, we propose an efficient one-shot search method to find the optimal path aggregation architecture, that is, we first train a super-net and then find the optimal candidate with an evolutionary algorithm. Experimental results demonstrate the efficacy of the proposed OPANAS for object detection: (1) OPANAS is more efficient than state-of-the-art methods (e.g., NAS-FPN and Auto-FPN), at significantly smaller searching cost (e.g., only 4 GPU days on MS-COCO); (2) the optimal architecture found by OPANAS significantly improves main-stream detectors including RetinaNet, Faster R-CNN and Cascade R-CNN, by 2.3 3.2 % mAP comparing to their FPN counterparts; and (3) a new state-of-the-art accuracy-speed trade-off (52.2 % mAP at 7.6 FPS) at smaller training costs than comparable state-of-the-arts.
****
Bridge To Answer: Structure-Aware Graph Interaction Network for Video Question Answering
Jungin Park, Jiyoung Lee, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15526-15535
This paper presents a novel method, termed Bridge to Answer, to infer correct answers for questions about a given video by leveraging adequate graph interactions of heterogeneous crossmodal graphs. To realize this, we learn question conditioned visual graphs by exploiting the relation between video and question to enable each visual node using question-to-visual interactions to encompass both visual and linguistic cues. In addition, we propose bridged visual-to-visual interactions to incorporate two complementary visual information on appearance and motion by placing the question graph as an intermediate bridge. This bridged architecture allows reliable message passing through compositional semantics of the question to generate an appropriate answer. As a result, our method can learn the question conditioned visual representations attributed to appearance and motion that show powerful capability for video question answering. Extensive experiments prove that the proposed method provides effective and superior performance than state-of-the-art methods on several benchmarks.
****
Learning Compositional Radiance Fields of Dynamic Human Heads
Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, Michael Zollhofer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5704-5713
Photorealistic rendering of dynamic humans is an important ability for telepresence systems, virtual shopping, synthetic data generation, and more. Recently, neural rendering methods, which combine techniques from computer graphics and machine learning, have created high-fidelity models of humans and objects. Some of these methods do not produce results with high-enough fidelity for driveable human models (Neural Volumes) whereas others have extremely long rendering times (Ne

RF). We propose a novel compositional 3D representation that combines the best of previous methods to produce both higher-resolution and faster results. Our representation bridges the gap between discrete and continuous volumetric representations by combining a coarse 3D-structure-aware grid of animation codes with a continuous learned scene function that maps every position and its corresponding local animation code to its view-dependent emitted radiance and local volume density. Differentiable volume rendering is employed to compute photo-realistic novel views of the human head and upper body as well as to train our novel representation end-to-end using only 2D supervision. In addition, we show that the learned dynamic radiance field can be used to synthesize novel unseen expressions based on a global animation code. Our approach achieves state-of-the-art results for synthesizing novel views of dynamic human heads and the upper body. See our project page for more results.
****
Partial Person Re-Identification With Part-Part Correspondence Learning
Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9105-9115
Driven by the success of deep learning, the last decade has seen rapid advances in person re-identification (re-ID). Nonetheless, most of approaches assume that the input is given with the fulfillment of expectations, while imperfect input remains rarely explored to date, which is a non-trivial problem since directly apply existing methods without adjustment can cause significant performance degradation. In this paper, we focus on recognizing partial (flawed) input with the assistance of proposed Part-Part Correspondence Learning (PPCL), a self-supervised learning framework that learns correspondence between image patches without any additional part-level supervision. Accordingly, we propose Part-Part Cycle (PP-Cycle) constraint and Part-Part Triplet (PP-Triplet) constraint that exploit the duality and uniqueness between corresponding image patches respectively. We verify our proposed PPCL on several partial person re-ID benchmarks. Experimental results demonstrate that our approach can surpass previous methods in terms of the standard evaluation metric.
****
Monte Carlo Scene Search for 3D Scene Understanding
Shreyas Hampali, Sinisa Stekovic, Sayan Deb Sarkar, Chetan S. Kumar, Friedrich Fraundorfer, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13804-13813
We explore how a general AI algorithm can be used for 3D scene understanding to reduce the need for training data. More exactly, we propose a modification of the Monte Carlo Tree Search (MCTS) algorithm to retrieve objects and room layouts from noisy RGB-D scans. While MCTS was developed as a game-playing algorithm, we show it can also be used for complex perception problems. Our adapted MCTS algorithm has few easy-to-tune hyperparameters and can optimise general losses. We use it to optimise the posterior probability of objects and room layout hypotheses given the RGB-D data. This results in an analysis-by-synthesis approach that explores the solution space by rendering the current solution and comparing it to the RGB-D observations. To perform this exploration even more efficiently, we propose simple changes to the standard MCTS' tree construction and exploration policy. We demonstrate our approach on the ScanNet dataset. Our method often retrieves configurations that are better than some manual annotations, especially on layouts.
****
Coarse-To-Fine Person Re-Identification With Auxiliary-Domain Classification and Second-Order Information Bottleneck
Anguo Zhang, Yueming Gao, Yuzhen Niu, Wenxi Liu, Yongcheng Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 598-607
Person re-identification (Re-ID) is to retrieve a particular person captured by different cameras, which is of great significance for security surveillance and pedestrian behavior analysis. However, due to the large intra-class variation of

a person across cameras, e.g., occlusions, illuminations, viewpoints, and poses, Re-ID is still a challenging task in the field of computer vision. In this paper, to attack the issues concerning with intra-class variation, we propose a coarse-to-fine Re-ID framework with the incorporation of auxiliary-domain classification (ADC) and second-order information bottleneck (2O-IB). In particular, as an auxiliary task, ADC is introduced to extract the coarse-grained essential features to distinguish a person from miscellaneous backgrounds, which leads to the effective coarse- and fine-grained feature representations for Re-ID. On the other hand, to cope with the redundancy, irrelevance, and noise contained in the Re-ID features caused by intra-class variations, we integrate 2O-IB into the network to compress and optimize the features, without increasing additional computation overhead during inference. Experimental results demonstrate that our proposed method significantly reduces the neural network output variance of intra-class person images and achieves the superior performance to state-of-the-art methods.
****
Transformer Tracking
Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8126-8135
Correlation acts as a critical role in the tracking field, especially in recent popular Siamese-based trackers. The correlation operation is a simple fusion manner to consider the similarity between the template and the search region. However, the correlation operation itself is a local linear matching process, leading to lose semantic information and fall into local optimum easily, which may be the bottleneck of designing high-accuracy tracking algorithms. Is there any better feature fusion method than correlation? To address this issue, inspired by Transformer, this work presents a novel attention-based feature fusion network, which effectively combines the template and search region features solely using attention. Specifically, the proposed method includes an ego-context augment module based on self-attention and a cross-feature augment module based on cross-attention. Finally, we present a Transformer tracking (named TransT) method based on the Siamese-like feature extraction backbone, the designed attention-based fusion mechanism, and the classification and regression head. Experiments show that our TransT achieves very promising results on six challenging datasets, especially on large-scale LaSOT, TrackingNet, and GOT-10k benchmarks. Our tracker runs at approximatively 50 fps on GPU. Code and models are available at https://github.com/chenxin-dlut/TransT.
****
Structured Multi-Level Interaction Network for Video Moment Localization via Language Query
Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7026-7035
We address the problem of localizing a specific moment described by a natural language query. Existing works interact the query with either video frame or moment proposal, and neglect the inherent structure of moment construction for both cross-modal understanding and video content comprehension, which are the two crucial challenges for this task. In this paper, we disentangle the activity moment into boundary and content. Based on the explored moment structure, we propose a novel Structured Multi-level Interaction Network (SMIN) to tackle this problem through multi-levels of cross-modal interaction coupled with content-boundary-moment interaction. In particular, for cross-modal interaction, we interact the sentence-level query with the whole moment while interact the word-level query with content and boundary, as in a coarse-to-fine manner. For content-boundary-moment interaction, we capture the insightful relations between boundary, content, and the whole moment proposal. Through multi-level interactions, the model obtains robust cross-modal representation for accurate moment localization. Extensive experiments conducted on three benchmarks (i.e., Charades-STA, ActivityNet-Captions, and TACoS) demonstrate the proposed approach outperforms the state-of-the-ar

t methods.
****

Structured Scene Memory for Vision-Language Navigation
Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8455-8464
Recently, numerous algorithms have been developed to tackle the problem of vision-language navigation (VLN), i.e., entailing an agent to navigate 3D environments through following linguistic instructions. However, current VLN agents simply store their past experiences/observations as latent states in recurrent networks, failing to capture environment layouts and make long-term planning. To address these limitations, we propose a crucial architecture, called Structured Scene Memory (SSM). It is compartmentalized enough to accurately memorize the percepts during navigation. It also serves as a structured scene representation, which captures and disentangles visual and geometric cues in the environment. SSM has a collect-read controller that adaptively collects information for supporting current decision making and mimics iterative algorithms for long-range reasoning. As SSM provides a complete action space, i.e., all the navigable places on the map, a frontier-exploration based navigation decision making strategy is introduced to enable efficient and global planning. Experiment results on two VLN datasets (i.e., R2R and R4R) show that our method achieves state-of-the-art performance on several metrics.
****

Unsupervised Pre-Training for Person Re-Identification
Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, Dong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14750-14759
In this paper, we present a large scale unlabeled person re-identification (Re-ID) dataset "LUPerson" and make the first attempt of performing unsupervised pre-training for improving the generalization ability of the learned person Re-ID feature representation. This is to address the problem that all existing person Re-ID datasets are all of limited scale due to the costly effort required for data annotation. Previous research tries to leverage models pre-trained on ImageNet to mitigate the shortage of person Re-ID data but suffers from the large domain gap between ImageNet and person Re-ID data. LUPerson is an unlabeled dataset of 4M images of over 200K identities, which is 30xlarger than the largest existing Re-ID dataset. It also covers a much diverse range of capturing environments (e.g., camera settings, scenes, etc.). Based on this dataset, we systematically study the key factors for learning Re-ID features from two perspectives: data augmentation and contrastive loss. Unsupervised pre-training performed on this large-scale dataset effectively leads to a generic Re-ID feature that can benefit all existing person Re-ID methods. Using our pre-trained model in some basic frameworks, our methods achieve state-of-the-art results without bells and whistles on four widely used Re-ID datasets: CUHK03, Market1501, DukeMTMC, and MSMT17. Our results also show that the performance improvement is more significant on small-scale target datasets or under few-shot setting.
****

Progressive Stage-Wise Learning for Unsupervised Feature Representation Enhancement
Zefan Li, Chenxi Liu, Alan Yuille, Bingbing Ni, Wenjun Zhang, Wen Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9767-9776
Unsupervised learning methods have recently shown their competitiveness against supervised training. Typically, these methods use a single objective to train the entire network. But one distinct advantage of unsupervised over supervised learning is that the former possesses more variety and freedom in designing the objective. In this work, we explore new dimensions of unsupervised learning by proposing the Progressive Stage-wise Learning (PSL) framework. For a given unsupervised task, we design multi-level tasks and define different learning stages for the deep network. Early learning stages are forced to focus on low-level tasks wh

ile late stages are guided to extract deeper information through harder tasks. We discover that by progressive stage-wise learning, unsupervised feature representation can be effectively enhanced. Our extensive experiments show that PSL consistently improves results for the leading unsupervised learning methods.
****

Domain-Specific Suppression for Adaptive Object Detection
Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, Yangyang Xia, Xishan Zhang, Shaoli Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9603-9612
Domain adaptation methods face performance degradation in object detection, as the complexity of tasks require more about the transferability of the model. We propose a new perspective on how CNN models gain the transferability, viewing the weights of a model as a series of motion patterns. The directions of weights, and the gradients, can be divided into domain-specific and domain-invariant parts, and the goal of domain adaptation is to concentrate on the domain-invariant direction while eliminating the disturbance from domain-specific one. Current UDA object detection methods view the two directions as a whole while optimizing, which will cause domain-invariant direction mismatch even if the output features are perfectly aligned. In this paper, we propose the domain-specific suppression, an exemplary and generalizable constraint to the original convolution gradients in backpropagation to detach the two parts of directions and suppress the domain-specific one. We further validate our theoretical analysis and methods on several domain adaptive object detection tasks, including weather, camera configuration, and synthetic to real-world adaptation. Our experiment results show significant advance over the state-of-the-art methods in the UDA object detection field, performing a promotion of 10.2 12.2% mAP on all these domain adaptation scenarios.
****

Few-Shot Object Detection via Classification Refinement and Distractor Retreatment
Yiting Li, Haiyue Zhu, Yu Cheng, Wenxin Wang, Chek Sing Teo, Cheng Xiang, Prahlad Vadakkepat, Tong Heng Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15395-15403
We aim to tackle the challenging Few-Shot Object Detection (FSOD) where data-scarce categories are presented during the model learning. The failure modes of FSOD are investigated that the performance degradation is mainly due to the classification incapability (false positives), which motivates us to address it from a novel aspect of hard example mining. Specifically, to address the intrinsic architecture limitation of common detectors under low-data constraint, we introduce a novel few-shot classification refinement mechanism where a decoupled Few-Shot Classification Network (FSCN) is employed to improve the classification. Moreover, we specially probe a commonly-overlooked but destructive issue of FSOD, i.e., the presence of distractor samples due to the incomplete annotations where images from base set may contain novel-class objects but remain unlabelled. Retreatment solutions are developed to eliminate the incurred false positives. For FSCN training, the distractor is formulated as a semi-supervised problem, where a distractor utilization loss is proposed to make proper use of it for boosting the data-scarce classes; while a Self-Supervised Dataset Pruning (SSDP) technique is developed to facilitate the few-shot adaptation of base detector. Experiments demonstrate that our proposed framework achieves the state-of-the-art FSOD performance on public datasets, e.g., Pascal VOC and MS-COCO.
****

D2IM-Net: Learning Detail Disentangled Implicit Fields From Single Images
Manyi Li, Hao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10246-10255
We present the first single-view 3D reconstruction network aimed at recovering geometric details from an input image which encompass both topological shape structures and surface features. Our key idea is to train the network to learn a detail disentangled reconstruction consisting of two functions, one implicit field representing the coarse 3D shape and the other capturing the details. Given an i

nput image, our network, coined D^2IM_Net, encodes it into global and local features which are respectively fed into two decoders. The base decoder uses the global features to reconstruct a coarse implicit field, while the detail decoder reconstructs, from the local features, two displacement maps, defined over the front and back sides of the captured object. The final 3D reconstruction is a fusion between the base shape and the displacement maps, with three losses enforcing the recovery of coarse shape, overall structure, and surface details via a novel Laplacian term.
****
Not Just Compete, but Collaborate: Local Image-to-Image Translation via Cooperative Mask Prediction
Daejin Kim, Mohammad Azam Khan, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6509-6518
Facial attribute editing aims to manipulate the image with the desired attribute while preserving the other details. Recently, generative adversarial networks along with the encoder-decoder architecture have been utilized for this task owing to their ability to create realistic images. However, the existing methods for the unpaired dataset cannot still preserve the attribute-irrelevant regions properly due to the absence of the ground truth image. This work proposes a novel, intuitive loss function called the CAM-consistency loss, which improves the consistency of an input image in image translation. While the existing cycle-consistency loss ensures that the image can be translated back, our approach makes the model further preserve the attribute-irrelevant regions even in a single translation to another domain by using the Grad-CAM output computed from the discriminator. Our CAM-consistency loss directly optimizes such a Grad-CAM output from the discriminator during training, in order to properly capture which local regions the generator should change while keeping the other regions unchanged. In this manner, our approach allows the generator and the discriminator to collaborate with each other to improve the image translation quality. In our experiments, we validate the effectiveness and versatility of our proposed CAM-consistency loss by applying it to several representative models for facial image editing, such as StarGAN, AttGAN, and STGAN.
****
Behavior-Driven Synthesis of Human Dynamics
Andreas Blattmann, Timo Milbich, Michael Dorkenwald, Bjorn Ommer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12236-12246
Generating and representing human behavior are of major importance for various computer vision applications. Commonly, human video synthesis represents behavior as sequences of postures while directly predicting their likely progressions or merely changing the appearance of the depicted persons, thus not being able to exercise control over their actual behavior during the synthesis process. In contrast, controlled behavior synthesis and transfer across individuals requires a deep understanding of body dynamics and calls for a representation of behavior that is independent of appearance and also of specific postures. In this work, we present a model for human behavior synthesis which learns a dedicated representation of human dynamics independent of postures. Using this representation, we are able to change the behavior of a person depicted in an arbitrary posture, or to even directly transfer behavior observed in a given video sequence. To this end, we propose a conditional variational framework which explicitly disentangles posture from behavior. We demonstrate the effectiveness of our approach on this novel task, evaluating capturing, transferring, and sampling fine-grained, diverse behavior, both quantitatively and qualitatively. Project page is available at https://cutt.ly/5l7rXEp
****
GAIA: A Transfer Learning System of Object Detection That Fits Your Needs
Xingyuan Bu, Junran Peng, Junjie Yan, Tieniu Tan, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 274-283
Transfer learning with pre-training on large-scale datasets has played an increa

singly significant role in computer vision and natural language processing recen
tly. However, as there exist numerous application scenarios that have distinctiv
e demands such as certain latency constraints and specialized data distributions
, it is prohibitively expensive to take advantage of large-scale pre-training fo
r per-task requirements. In this paper, we focus on the area of object detection
 and present a transfer learning system named GAIA, which could automatically an
d efficiently give birth to customized solutions according to heterogeneous down
stream needs. GAIA is capable of providing powerful pre-trained weights, selecti
ng models that conform to downstream demands such as latency constraints and spe
cified data domains, and collecting relevant data for practitioners who have ver
y few datapoints for their tasks. With GAIA, we achieve promising results on COC
O, Objects365, Open Images, Caltech, CityPersons, and UODB which is a collection
 of datasets including KITTI, VOC, WiderFace, DOTA, Clipart, Comic, and more. Ta
king COCO as an example, GAIA is able to efficiently produce models covering a w
ide range of latency from 16ms to 53ms, and yields AP from 38.2 to 46.5 without
whistles and bells. To benefit every practitioner in the community of object det
ection, we would release our pre-trained models and code.
****
IronMask: Modular Architecture for Protecting Deep Face Template
Sunpill Kim, Yunseong Jeong, Jinsu Kim, Jungkon Kim, Hyung Tae Lee, Jae Hong Seo
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2021, pp. 16125-16134
Convolutional neural networks have made remarkable progress in the face recognit
ion field. The more the technology of face recognition advances, the greater dis
criminative features into a face template. However, this increases the threat to
 user privacy in case the template is exposed. In this paper, we present a modul
ar architecture for face template protection, called IronMask, that can be combi
ned with any face recognition system using angular distance metric. We circumven
t the need for binarization, which is the main cause of performance degradation
in most existing face template protections, by proposing a new real-valued error
-correcting-code that is compatible with real-valued templates and can therefore
, minimize performance degradation. We evaluate the efficacy of IronMask by exte
nsive experiments on two face recognitions, ArcFace and CosFace with three datas
ets, CMU-Multi-PIE, FEI, and Color-FERET. According to our experimental results,
 IronMask achieves a true accept rate (TAR) of 99.79% at a false accept rate (FA
R) of 0.0005% when combined with ArcFace, and 95.78% TAR at 0% FAR with CosFace,
 while providing at least 115-bit security against known attacks.
****
Learning To Recommend Frame for Interactive Video Object Segmentation in the Wil
d
Zhaoyuan Yin, Jia Zheng, Weixin Luo, Shenhan Qian, Hanling Zhang, Shenghua Gao;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2021, pp. 15445-15454
This paper proposes a framework for the interactive video object segmentation (V
OS) in the wild where users can choose some frames for annotations iteratively.
Then, based on the user annotations, a segmentation algorithm refines the masks.
 The previous interactive VOS paradigm selects the frame with some worst evaluat
ion metric, and the ground truth is required for calculating the evaluation metr
ic, which is impractical in the testing phase. In contrast, in this paper, we ad
vocate that the frame with the worst evaluation metric may not be exactly the mo
st valuable frame that leads to the most performance improvement across the vide
o. Thus, we formulate the frame selection problem in the interactive VOS as a Ma
rkov Decision Process, where an agent is learned to recommend the frame under a
deep reinforcement learning framework. The learned agent can automatically deter
mine the most valuable frame, making the interactive setting more practical in t
he wild. Experimental results on the public datasets show the effectiveness of o
ur learned agent without any changes to the underlying VOS algorithms. Our data,
 code, and models are available at https://github.com/svip-lab/IVOS-W.
****
DSRNA: Differentiable Search of Robust Neural Architectures

Ramtin Hosseini, Xingyi Yang, Pengtao Xie; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6196-6205
In deep learning applications, the architectures of deep neural networks are cru
cial in achieving high accuracy. Many methods have been proposed to search for h
igh-performance neural architectures automatically. However, these searched arch
itectures are prone to adversarial attacks. A small perturbation of the input da
ta can render the architecture to change prediction outcomes significantly. To a
ddress this problem, we propose methods to perform differentiable searches of ro
bust neural architectures. In our methods, two differentiable metrics are define
d to measure architectures' robustness, based on certified lower bound and Jacob
ian norm bound. Then we search for robust architectures by maximizing the robust
ness metrics. Different from previous approaches which aim to improve architectu
res' robustness in an implicit way: performing adversarial training and injectin
g random noise, our methods explicitly and directly maximize robustness metrics
to harvest robust architectures. On CIFAR-10, ImageNet, and MNIST, we perform ga
me-based evaluation and verification-based evaluation on the robustness of our m
ethods. The experimental results show that our methods 1) are more robust to var
ious norm-bound attacks than several robust NAS baselines; 2) are more accurate
than baselines when there are no attacks; 3) have significantly higher certified
 lower bounds than baselines.
****
Reconstructing 3D Human Pose by Watching Humans in the Mirror
Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, Xiaowei Zhou; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
12814-12823
In this paper, we introduce the new task of reconstructing 3D human pose from a
single image in which we can see the person and the person's image through a mir
ror. Compared to general scenarios of 3D pose estimation from a single view, the
 mirror reflection provides an additional view for resolving the depth ambiguity
. We develop an optimization-based approach that exploits mirror symmetry constr
aints for accurate 3D pose reconstruction. We also provide a method to estimate
the surface normal of the mirror from vanishing points in the single image. To v
alidate the proposed approach, we collect a large-scale dataset named Mirrored-H
uman, which covers a large variety of human subjects, poses and backgrounds. The
 experiments demonstrate that, when trained on Mirrored-Human with our reconstru
cted 3D poses as pseudo ground-truth, the accuracy and generalizability of exist
ing single-view 3D pose estimators can be largely improved.
****
Spk2ImgNet: Learning To Reconstruct Dynamic Scene From Continuous Spike Stream
Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, Tiejun Huang; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
pp. 11996-12005
The recently invented retina-inspired spike camera has shown great potential for
 capturing dynamic scenes. Different from the conventional digital cameras that
compact the photoelectric information within the exposure interval into a single
 snapshot, the spike camera produces a continuous spike stream to record the dyn
amic light intensity variation process. For spike cameras, image reconstruction
remains an important and challenging issue. To this end, this paper develops a s
pike-to-image neural network (Spk2ImgNet) to reconstruct the dynamic scene from
the continuous spike stream. In particular, to handle the challenges brought by
both noise and high-speed motion, we propose a hierarchical architecture to expl
oit the temporal correlation of the spike stream progressively. Firstly, a spati
ally adaptive light inference subnet is proposed to exploit the local temporal c
orrelation, producing basic light intensity estimates of different moments. Then
, a pyramid deformable alignment is utilized to align the intermediate features
such that the feature fusion module can exploit the long-term temporal correlati
on, while avoiding undesired motion blur. In addition, to train the network, we
simulate the working mechanism of spike camera to generate a large-scale spike d
ataset composed of spike streams and corresponding ground truth images. Experime
ntal results demonstrate that the proposed network evidently outperforms the sta

te-of-the-art spike camera reconstruction methods.
****
MonoRUn: Monocular 3D Object Detection by Reconstruction and Uncertainty Propaga tion
Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, Lu Xiong; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1 0379-10388
Object localization in 3D space is a challenging aspect in monocular 3D object d etection. Recent advances in 6DoF pose estimation have shown that predicting den se 2D-3D correspondence maps between image and object 3D model and then estimati ng object pose via Perspective-n-Point (PnP) algorithm can achieve remarkable lo calization accuracy. Yet these methods rely on training with ground truth of obj ect geometry, which is difficult to acquire in real outdoor scenes. To address t his issue, we propose MonoRUn, a novel detection framework that learns dense cor respondences and geometry in a self-supervised manner, with simple 3D bounding b ox annotations. To regress the pixel-related 3D object coordinates, we employ a regional reconstruction network with uncertainty awareness. For self-supervised training, the predicted 3D coordinates are projected back to the image plane. A Robust KL loss is proposed to minimize the uncertainty-weighted reprojection err or. During testing phase, we exploit the network uncertainty by propagating it t hrough all downstream modules. More specifically, the uncertainty-driven PnP alg orithm is leveraged to estimate object pose and its covariance. Extensive experi ments demonstrate that our proposed approach outperforms current state-of-the-ar t methods on KITTI benchmark.
****
Complete & Label: A Domain Adaptation Approach to Semantic Segmentation of LiDAR Point Clouds
Li Yi, Boqing Gong, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15363-15373
We study an unsupervised domain adaptation problem for the semantic labeling of 3D point clouds, with a particular focus on domain discrepancies induced by diff erent LiDAR sensors. Based on the observation that sparse 3D point clouds are sa mpled from 3D surfaces, we take a Complete and Label approach to recover the und erlying surfaces before passing them to a segmentation network. Specifically, we design a Sparse Voxel Completion Network (SVCN) to complete the 3D surfaces of a sparse point cloud. Unlike semantic labels, to obtain training pairs for SVCN requires no manual labeling. We also introduce local adversarial learning to mod el the surface prior. The recovered 3D surfaces serve as a canonical domain, fro m which semantic labels can transfer across different LiDAR sensors. Experiments and ablation studies with our new benchmark for cross-domain semantic labeling of LiDAR data show that the proposed approach provides 6.3-37.6% better performa nce than previous domain adaptation methods.
****
GMOT-40: A Benchmark for Generic Multiple Object Tracking
Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, Haibin Ling; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2021, pp. 6719-6728
Multiple Object Tracking (MOT) has witnessed remarkable advances in recent years . However, existing studies dominantly request prior knowledge of the tracking t arget (eg, pedestrians), and hence may not generalize well to unseen categories. In contrast, Generic Multiple Object Tracking (GMOT), which requires little pri or information about the target, is largely under-explored. In this paper, we ma ke contributions to boost the study of GMOT in three aspects. First, we construc t the first publicly available dense GMOT dataset, dubbed GMOT-40, which contain s 40 carefully annotated sequences evenly distributed among 10 object categories . In addition, two tracking protocols are adopted to evaluate different characte ristics of tracking algorithms. Second, by noting the lack of devoted tracking a lgorithms, we have designed a series of baseline GMOT algorithms. Third, we perf orm thorough evaluations on GMOT-40, involving popular MOT algorithms (with nece ssary modifications) and the proposed baselines. The GMOT-40 benchmark is public

ly available at https://github.com/Spritea/GMOT40.
****
Few-Shot Image Generation via Cross-Domain Correspondence
Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, Richard Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10743-10752
Training generative models, such as GANs, on a target domain containing limited examples (e.g., 10) can easily result in overfitting. In this work, we seek to utilize a large source domain for pretraining and transfer the diversity information from source to target. We propose to preserve the relative similarities and differences between instances in the source via a novel cross-domain distance consistency loss. To further reduce overfitting, we present an anchor-based strategy to encourage different levels of realism over different regions in the latent space. With extensive results in both photorealistic and non-photorealistic domains, we demonstrate qualitatively and quantitatively that our few-shot model automatically discovers correspondences between source and target domains and generates more diverse and realistic images than previous methods.
****
Hierarchical Lovasz Embeddings for Proposal-Free Panoptic Segmentation
Tommi Kerola, Jie Li, Atsushi Kanehira, Yasunori Kudo, Alexis Vallet, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14413-14423
Panoptic segmentation brings together two separate tasks: instance and semantic segmentation. Although they are related, unifying them faces an apparent paradox: how to learn simultaneously instance-specific and category-specific (i.e. instance-agnostic) representations jointly. Hence, state-of-the-art panoptic segmentation methods use complex models with a distinct stream for each task. In contrast, we propose Hierarchical Lovasz Embeddings, per pixel feature vectors that simultaneously encode instance- and category-level discriminative information. We use a hierarchical Lovasz hinge loss to learn a low-dimensional embedding space structured into a unified semantic and instance hierarchy without requiring separate network branches or object proposals. Besides modeling instances precisely in a proposal-free manner, our Hierarchical Lovasz Embeddings generalize to categories by using a simple Nearest-Class-Mean classifier, including for non-instance ""stuff"" classes where instance segmentation methods are not applicable. Our simple model achieves state-of-the-art results compared to existing proposal-free panoptic segmentation methods on Cityscapes, COCO, and Mapillary Vistas. Furthermore, our model demonstrates temporal stability between video frames.
****
Neural Body: Implicit Neural Representations With Structured Latent Codes for Novel View Synthesis of Dynamic Humans
Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9054-9063
This paper addresses the challenge of novel view synthesis for a human performer from a very sparse set of camera views. Some recent works have shown that learning implicit neural representations of 3D scenes achieves remarkable view synthesis quality given dense input views. However, the representation learning will be ill-posed if the views are highly sparse. To solve this ill-posed problem, our key idea is to integrate observations over video frames. To this end, we propose Neural Body, a new human body representation which assumes that the learned neural representations at different frames share the same set of latent codes anchored to a deformable mesh, so that the observations across frames can be naturally integrated. The deformable mesh also provides geometric guidance for the network to learn 3D representations more efficiently. Experiments on a newly collected multi-view dataset show that our approach outperforms prior works by a large margin in terms of the novel view synthesis quality. We also demonstrate the capability of our approach to reconstruct a moving person from a monocular video on the People-Snapshot dataset. We will release the code and dataset for reproducibility.

****
Cross-Modal Collaborative Representation Learning and a Large-Scale RGBT Benchmark for Crowd Counting
Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4823-4833
Crowd counting is a fundamental yet challenging task, which desires rich information to generate pixel-wise crowd density maps. However, most previous methods only used the limited information of RGB images and cannot well discover potential pedestrians in unconstrained scenarios. In this work, we find that incorporating optical and thermal information can greatly help to recognize pedestrians. To promote future researches in this field, we introduce a large-scale RGBT Crowd Counting (RGBT-CC) benchmark, which contains 2,030 pairs of RGB-thermal images with 138,389 annotated people. Furthermore, to facilitate the multimodal crowd counting, we propose a cross-modal collaborative representation learning framework, which consists of multiple modality-specific branches, a modality-shared branch, and an Information Aggregation-Distribution Module (IADM) to capture the complementary information of different modalities fully. Specifically, our IADM incorporates two collaborative information transfers to dynamically enhance the modality-shared and modality-specific representations with a dual information propagation mechanism. Extensive experiments conducted on the RGBT-CC benchmark demonstrate the effectiveness of our framework for RGBT crowd counting. Moreover, the proposed approach is universal for multimodal crowd counting and is also capable to achieve superior performance on the ShanghaiTechRGBD dataset. Finally, our source code and benchmark have been released at http://lingboliu.com/RGBT_Crowd_Counting.html.
****
Weakly Supervised Video Salient Object Detection
Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16826-16835
Significant performance improvement has been achieved for fully-supervised video salient object detection with the pixel-wise labeled training datasets, which are timeconsuming and expensive to obtain. To relieve the burden of data annotation, we present the first weakly supervised video salient object detection model based on relabeled "fixation guided scribble annotations". Specifically, an "Appearance-motion fusion module" and bidirectional ConvLSTM based framework are proposed to achieve effective multi-modal learning and long-term temporal context modeling based on our new weak annotations. Further, we design a novel foreground-background similarity loss to further explore the labeling similarity across frames. A weak annotation boosting strategy is also introduced to boost our model performance with a new pseudo-label generation technique. Extensive experimental results on six benchmark video saliency detection datasets illustrate the effectiveness of our solution.
****
Pixel-Wise Anomaly Detection in Complex Driving Scenes
Giancarlo Di Biase, Hermann Blum, Roland Siegwart, Cesar Cadena; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16918-16927
The inability of state-of-the-art semantic segmentation methods to detect anomaly instances hinders them from being deployed in safety-critical and complex applications, such as autonomous driving. Recent approaches have focused on either leveraging segmentation uncertainty to identify anomalous areas or re-synthesizing the image from the semantic label map to find dissimilarities with the input image. In this work, we demonstrate that these two methodologies contain complementary information and can be combined to produce robust predictions for anomaly segmentation. We present a pixel-wise anomaly detection framework that uses uncertainty maps to improve over existing re-synthesis methods in finding dissimilarities between the input and generated images. Our approach works as a general framework around already trained segmentation networks, which ensures anomaly dete

ction without compromising segmentation accuracy, while significantly outperform ing all similar methods. Top-2 performance across a range of different anomaly d atasets shows the robustness of our approach to handling different anomaly insta nces.
****
Learning To Associate Every Segment for Video Panoptic Segmentation
Sanghyun Woo, Dahun Kim, Joon-Young Lee, In So Kweon; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2705-2 714
Temporal correspondence -- linking pixels or objects across frames -- is a funda mental supervisory signal for the video models. For the panoptic understanding o f dynamic scenes, we further extend this concept to every segment. Specifically, we aim to learn coarse segment-level matching and fine pixel-level matching tog ether. We implement this idea by designing two novel learning objectives. To val idate our proposals, we adopt a deep siamese model and train the model to learn the temporal correspondence on two different levels (i.e., segment and pixel) al ong with the target task. At inference time, the model processes each frame inde pendently without any extra computation and post-processing. We show that our pe r-frame inference model can achieve new state-of-the-art results on Cityscapes-V PS and VIPER datasets. Moreover, due to its high efficiency, the model runs in a fraction of time (3x) compared to the previous state-of-the-art approach. The c odes and models will be released.
****
Variational Transformer Networks for Layout Generation
Diego Martin Arroyo, Janis Postels, Federico Tombari; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13642- 13652
Generative models able to synthesize layouts of different kinds (e.g. documents, user interfaces or furniture arrangements) are a useful tool to aid design proc esses and as a first step in the generation of synthetic data, among other tasks . We exploit the properties of self-attention layers to capture high level relat ionships between elements in a layout, and use these as the building blocks of t he well-known Variational Autoencoder (VAE) formulation. Our proposed Variationa l Transformer Network (VTN) is capable of learning margins, alignments and other global design rules without explicit supervision. Layouts sampled from our mode l have a high degree of resemblance to the training data, while demonstrating ap pealing diversity. In an extensive evaluation on publicly available benchmarks f or different layout types VTNs achieve state-of-the-art diversity and perceptual quality. Additionally, we show the capabilities of this method as part of a doc ument layout detection pipeline.
****
Mitigating Face Recognition Bias via Group Adaptive Classifier
Sixue Gong, Xiaoming Liu, Anil K. Jain; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3414-3424
Face recognition is known to exhibit bias -- subjects in a certain demographic g roup can be better recognized than other groups. This work aims to learn a fair face representation, where faces of every group could be more equally represente d. Our proposed group adaptive classifier mitigates bias by using adaptive convo lution kernels and attention mechanisms on faces based on their demographic attr ibutes. The adaptive module comprises kernel masks and channel-wise attention ma ps for each demographic group so as to activate different facial regions for ide ntification, leading to more discriminative features pertinent to their demograp hics. Our introduced automated adaptation strategy determines whether to apply a daptation to a certain layer by iteratively computing the dissimilarity among de mographic-adaptive parameters. A new de-biasing loss function is proposed to mit igate the gap of average intra-class distance between demographic groups. Experi ments on face benchmarks (RFW, LFW, IJB-A, and IJB-C) show that our work is able to mitigate face recognition bias across demographic groups while maintaining t he competitive accuracy.
****

A Peek Into the Reasoning of Neural Networks: Interpreting With Structural Visua
l Concepts
Yunhao Ge, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laur
ent Itti, Ziyan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2021, pp. 2195-2204
Despite substantial progress in applying neural networks (NN) to a wide variety
of areas, they still largely suffer from a lack of transparency and interpretabi
lity. While recent developments in explainable artificial intelligence attempt t
o bridge this gap (e.g., by visualizing the correlation between input pixels and
 final outputs), these approaches are limited to explaining low-level relationsh
ips, and crucially, do not provide insights on error correction. In this work, w
e propose a framework (VRX) to interpret classification NNs with intuitive struc
tural visual concepts. Given a trained classification model, the proposed VRX ex
tracts relevant class-specific visual concepts and organizes them using structur
al concept graphs (SCG) based on pairwise concept relationships. By means of kno
wledge distillation, we show VRX can take a step towards mimicking the reasoning
 process of NNs and provide logical, concept-level explanations for final model
decisions. With extensive experiments, we empirically show VRX can meaningfully
answer "why" and "why not" questions about the prediction, providing easy-to-und
erstand insights about the reasoning process. We also show that these insights c
an potentially provide guidance on improving NN's performance.
****
Three Birds with One Stone: Multi-Task Temporal Action Detection via Recycling T
emporal Annotations
Zhihui Li, Lina Yao; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2021, pp. 4751-4760
Temporal action detection on unconstrained videos has seen significant research
progress in recent years. Deep learning has achieved enormous success in this di
rection. However, collecting large-scale temporal detection datasets to ensuring
 promising performance in the real-world is a laborious, impractical and time co
nsuming process. Accordingly, we present a novel improved temporal action locali
zation model that is better able to take advantage of limited labeled data avail
able. Specifically, we design two auxiliary tasks by reconstructing the availabl
e label information and then facilitate the learning of the temporal action dete
ction model. Each task generates their supervision signal by recycling the origi
nal annotations, and are jointly trained with the temporal action detection mode
l in a multi-task learning fashion. Note that the proposed approach can be plugg
able to any region proposal based temporal action detection models. We conduct e
xtensive experiments on three benchmark datasets, namely THUMOS'14, Charades and
 ActivityNet. Our experimental results confirm the effectiveness of the proposed
 model.
****
A Dual Iterative Refinement Method for Non-Rigid Shape Matching
Rui Xiang, Rongjie Lai, Hongkai Zhao; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15930-15939
In this work, a robust and efficient dual iterative refinement (DIR) method is p
roposed for dense correspondence between two nearly isometric shapes. The key id
ea is to use dual information, such as spatial and spectral, or local and global
 features, in a complementary and effective way, and extract more accurate infor
mation from current iteration to use for the next iteration. In each DIR iterati
on, starting from current correspondence, a zoom-in process at each point is use
d to select well matched anchor pairs by a local mapping distortion criterion. T
hese selected anchor pairs are then used to align spectral features (or other ap
propriate global features) whose dimension adaptively matches the capacity of th
e selected anchor pairs. Thanks to the effective combination of complementary in
formation in a data-adaptive way, DIR is not only efficient but also robust to r
ender accurate results within a few iterations. By choosing appropriate dual fea
tures, DIR has the flexibility to handle patch and partial matching as well. Ext
ensive experiments on various data sets demonstrate the superiority of DIR over
other state-of-the-art methods in terms of both accuracy and efficiency.

****
Image Super-Resolution With Non-Local Sparse Attention
Yiqun Mei, Yuchen Fan, Yuqian Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3517-3526
Both non-local (NL) operation and sparse representation are crucial for Single Image Super-Resolution (SISR). In this paper, we investigate their combinations and propose a novel Non-Local Sparse Attention (NLSA) with dynamic sparse attention pattern. NLSA is designed to retain long-range modeling capability from NL operation while enjoying robustness and high-efficiency of sparse representation. Specifically, NLSA rectifies NL attention with spherical locality sensitive hashing (LSH) that partitions the input space into hash buckets of related features. For every query signal, NLSA assigns a bucket to it and only computes attention within the bucket. The resulting sparse attention prevents the model from attending to locations that are noisy and less-informative, while reducing the computational cost from quadratic to asymptotic linear with respect to the spatial size. Extensive experiments validate the effectiveness and efficiency of NLSA. With a few non-local sparse attention modules, our architecture, called non-local sparse network (NLSN), reaches state-of-the-art performance for SISR quantitatively and qualitatively.
****
3D Video Stabilization With Depth Estimation by CNN-Based Optimization
Yao-Chih Lee, Kuan-Wei Tseng, Yu-Ta Chen, Chien-Cheng Chen, Chu-Song Chen, Yi-Ping Hung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10621-10630
Video stabilization is an essential component of visual quality enhancement. Early methods rely on feature tracking to recover either 2D or 3D frame motion, which suffer from the robustness of local feature extraction and tracking in shaky videos. Recently, learning-based methods seek to find frame transformations with high-level information via deep neural networks to overcome the robustness issue of feature tracking. Nevertheless, to our best knowledge, no learning-based methods leverage 3D cues for the transformation inference yet; hence they would lead to artifacts on complex scene-depth scenarios. In this paper, we propose Deep 3D Stabilizer, a novel 3D depth-based learning method for video stabilization. We take advantage of the recent self-supervised framework on jointly learning depth and camera ego-motion estimation on raw videos. Our approach requires no data for pre-training but stabilizes the input video via 3D reconstruction directly. The rectification stage incorporates the 3D scene depth and camera motion to smooth the camera trajectory and synthesize the stabilized video. Unlike most one-size-fits-all learning-based methods, our smoothing algorithm allows users to manipulate the stability of a video efficiently. Experimental results on challenging benchmarks show that the proposed solution consistently outperforms the state-of-the-art methods on almost all motion categories.
****
Predicting Human Scanpaths in Visual Question Answering
Xianyu Chen, Ming Jiang, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10876-10885
Attention has been an important mechanism for both humans and computer vision systems. While state-of-the-art models to predict attention focus on estimating a static probabilistic saliency map with free-viewing behavior, real-life scenarios are filled with tasks of varying types and complexities, and visual exploration is a temporal process that contributes to task performance. To bridge the gap, we conduct a first study to understand and predict the temporal sequences of eye fixations (a.k.a. scanpaths) during performing general tasks, and examine how scanpaths affect task performance. We present a new deep reinforcement learning method to predict scanpaths leading to different performances in visual question answering. Conditioned on a task guidance map, the proposed model learns question-specific attention patterns to generate scanpaths. It addresses the exposure bias in scanpath prediction with self-critical sequence training and designs a Consistency-Divergence loss to generate distinguishable scanpaths between correct and incorrect answers. The proposed model not only accurately predicts the spat

io-temporal patterns of human behavior in visual question answering, such as fixation position, duration, and order, but also generalizes to free-viewing and visual search tasks, achieving human-level performance in all tasks and significantly outperforming the state of the art.
****
DetectoRS: Detecting Objects With Recursive Feature Pyramid and Switchable Atrous Convolution
Siyuan Qiao, Liang-Chieh Chen, Alan Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10213-10224
Many modern object detectors demonstrate outstanding performances by using the mechanism of looking and thinking twice. In this paper, we explore this mechanism in the backbone design for object detection. At the macro level, we propose Recursive Feature Pyramid, which incorporates extra feedback connections from Feature Pyramid Networks into the bottom-up backbone layers. At the micro level, we propose Switchable Atrous Convolution, which convolves the features with different atrous rates and gathers the results using switch functions. Combining them results in DetectoRS, which significantly improves the performances of object detection. On COCO test-dev, DetectoRS achieves state-of-the-art 55.7% box AP for object detection, 48.5% mask AP for instance segmentation, and 50.0% PQ for panoptic segmentation. The code is made publicly available.
****
SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks
Shunsuke Saito, Jinlong Yang, Qianli Ma, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2886-2897
We present SCANimate, an end-to-end trainable framework that takes raw 3D scans of a clothed human and turns them into an animatable avatar. These avatars are driven by pose parameters and have realistic clothing that moves and deforms naturally. SCANimate does not rely on a customized mesh template or surface mesh registration. We observe that fitting a parametric 3D body model, like SMPL, to a clothed human scan is tractable while surface registration of the body topology to the scan is often not, because clothing can deviate significantly from the body shape. We also observe that articulated transformations are invertible, resulting in geometric cycle-consistency in the posed and unposed shapes. These observations lead us to a weakly supervised learning method that aligns scans into a canonical pose by disentangling articulated deformations without template-based surface registration. Furthermore, to complete missing regions in the aligned scans while modeling pose-dependent deformations, we introduce a locally pose-aware implicit function that learns to complete and model geometry with learned pose correctives. In contrast to commonly used global pose embeddings, our local pose conditioning significantly reduces long-range spurious correlations and improves generalization to unseen poses, especially when training data is limited. Our method can be applied to pose-aware appearance modeling to generate a fully textured avatar. We demonstrate our approach on various clothing types with different amounts of training data, outperforming existing solutions and other variants in terms of fidelity and generality in every setting. The code is available at https://scanimate.is.tue.mpg.de
****
Improving Accuracy of Binary Neural Networks Using Unbalanced Activation Distribution
Hyungjun Kim, Jihoon Park, Changhun Lee, Jae-Joon Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7862-7871
Binarization of neural network models is considered as one of the promising methods to deploy deep neural network models on resource-constrained environments such as mobile devices. However, Binary Neural Networks (BNNs) tend to suffer from severe accuracy degradation compared to the full-precision counterpart model. Several techniques were proposed to improve the accuracy of BNNs. One of the approaches is to balance the distribution of binary activations so that the amount of information in the binary activations becomes maximum. Based on extensive anal

ysis, in stark contrast to previous work, we argue that unbalanced activation distribution can actually improve the accuracy of BNNs. We also show that adjusting the threshold values of binary activation functions results in the unbalanced distribution of the binary activation, which increases the accuracy of BNN models. Experimental results show that the accuracy of previous BNN models (e.g. XNOR-Net and Bi-Real-Net) can be improved by simply shifting the threshold values of binary activation functions without requiring any other modification.
****
Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR Segmentation
Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9939-9948
State-of-the-art methods for large-scale driving-scene LiDAR segmentation often project the point clouds to 2D space and then process them via 2D convolution. Although this corporation shows the competitiveness in the point cloud, it inevitably alters and abandons the 3D topology and geometric relations. A natural remedy is to utilize the 3D voxelization and 3D convolution network. However, we found that in the outdoor point cloud, the improvement obtained in this way is quite limited. An important reason is the property of the outdoor point cloud, namely sparsity and varying density. Motivated by this investigation, we propose a new framework for the outdoor LiDAR segmentation, where cylindrical partition and asymmetrical 3D convolution networks are designed to explore the 3D geometric pattern while maintaining these inherent properties. Moreover, a point-wise refinement module is introduced to alleviate the interference of lossy voxel-based label encoding. We evaluate the proposed model on two large-scale datasets , i.e., SemanticKITTI and nuScenes. Our method achieves the 1st place in the leaderboard of SemanticKITTI and outperforms existing methods on nuScenes with a noticeable margin. Furthermore, the proposed 3D framework also generalizes well to LiDAR panoptic segmentation and LiDAR 3D detection.
****
SMPLicit: Topology-Aware Generative Model for Clothed People
Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11875-11885
In this paper we introduce SMPLicit, a novel generative model to jointly represent body pose, shape and clothing geometry. In contrast to existing learning-based approaches that require training specific models for each type of garment, SMPLicit can represent in a unified manner different garment topologies (e.g. from sleeveless tops to hoodies and to open jackets), while controlling other properties like the garment size or tightness/looseness. We show our model to be applicable to a large variety of garments including T-shirts, hoodies, jackets, shorts, pants, skirts, shoes and even hair. The representation flexibility of SMPLicit builds upon an implicit model conditioned with the SMPL human body parameters and a learnable latent space which is semantically interpretable and aligned with the clothing attributes. The proposed model is fully differentiable, allowing for its use into larger end-to-end trainable systems. In the experimental section, we demonstrate SMPLicit can be readily used for fitting 3D scans and for 3D reconstruction in images of dressed people. In both cases we are able to go beyond state of the art, by retrieving complex garment geometries, handling situations with multiple clothing layers and providing a tool for easy outfit editing. To stimulate further research in this direction, we will make our code and model publicly available at https://link/smplicit/.
****
Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization
Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J. Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, Ting Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12793-12802
We introduce a novel representation learning method to disentangle pose-dependen

t as well as view-dependent factors from 2D human poses. The method trains a network using cross-view mutual information maximization (CV-MIM) which maximizes mutual information of the same pose performed from different viewpoints in a contrastive learning manner. We further propose two regularization terms to ensure disentanglement and smoothness of the learned representations. The resulting pose representations can be used for cross-view action recognition. To evaluate the power of the learned representations, in addition to the conventional fully-supervised action recognition settings, we introduce a novel task called single-shot cross-view action recognition. This task trains models with actions from only one single viewpoint while models are evaluated on poses captured from all possible viewpoints. We evaluate the learned representations on standard benchmarks for action recognition, and show that (i) CV-MIM performs competitively compared with the state-of-the-art models in the fully-supervised scenarios; (ii) CV-MIM outperforms other competing methods by a large margin in the single-shot cross-view setting; (iii) and the learned representations can significantly boost the performance when reducing the amount of supervised training data. Our code is made publicly available at https://github.com/google-research/google-research/tree/master/poem.
****

Non-Salient Region Object Mining for Weakly Supervised Semantic Segmentation
Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2623-2632
Semantic segmentation aims to classify every pixel of an input image. Considering the difficulty of acquiring dense labels, researchers have recently been resorting to weak labels to alleviate the annotation burden of segmentation. However, existing works mainly concentrate on expanding the seed of pseudo labels within the image's salient region. In this work, we propose a non-salient region object mining approach for weakly supervised semantic segmentation. We introduce a graph-based global reasoning unit to strengthen the classification network's ability to capture global relations among disjoint and distant regions. This helps the network activate the object features outside the salient area. To further mine the non-salient region objects, we propose to exert the segmentation network's self-correction ability. Specifically, a potential object mining module is proposed to reduce the false-negative rate in pseudo labels. Moreover, we propose a non-salient region masking module for complex images to generate masked pseudo labels. Our non-salient region masking module helps further discover the objects in the non-salient region. Extensive experiments on the PASCAL VOC dataset demonstrate state-of-the-art results compared to current methods.
****

DCT-Mask: Discrete Cosine Transform Mask Representation for Instance Segmentation
Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, Kewei Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8720-8729
Binary grid mask representation is broadly used in instance segmentation. A representative instantiation is Mask R-CNN which predicts masks on a 28*28 binary grid. Generally, a low-resolution grid is not sufficient to capture the details, while a high-resolution grid dramatically increases the training complexity. In this paper, we propose a new mask representation by applying the discrete cosine transform(DCT) to encode the high-resolution binary grid mask into a compact vector. Our method, termed DCT-Mask, could be easily integrated into most pixel-based instance segmentation methods. Without any bells and whistles, DCT-Mask yields significant gains on different frameworks, backbones, datasets, and training schedules. It does not require any pre-processing or pre-training, and almost no harm to the running speed. Especially, for higher-quality annotations and more complex backbones, our method has a greater improvement. Moreover, we analyze the performance of our method from the perspective of the quality of mask representation. The main reason why DCT-Mask works well is that it obtains a high-quality mask representation with low complexity.

****
Bridging the Visual Gap: Wide-Range Image Blending
Chia-Ni Lu, Ya-Chu Chang, Wei-Chen Chiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 843-851
In this paper we propose a new problem scenario in image processing, wide-range image blending, which aims to smoothly merge two different input photos into a panorama by generating novel image content for the intermediate region between them. Although such problem is closely related to the topics of image inpainting, image outpainting, and image blending, none of the approaches from these topics is able to easily address it. We introduce an effective deep-learning model to realize wide-range image blending, where a novel Bidirectional Content Transfer module is proposed to perform the conditional prediction for the feature representation of the intermediate region via recurrent neural networks. In addition to ensuring the spatial and semantic consistency during the blending, we also adopt the contextual attention mechanism as well as the adversarial learning scheme in our proposed method for improving the visual quality of the resultant panorama. We experimentally demonstrate that our proposed method is not only able to produce visually appealing results for wide-range image blending, but also able to provide superior performance with respect to several baselines built upon the state-of-the-art image inpainting and outpainting approaches.
****
A Realistic Evaluation of Semi-Supervised Learning for Fine-Grained Classification
Jong-Chyi Su, Zezhou Cheng, Subhransu Maji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12966-12975
We evaluate the effectiveness of semi-supervised learning (SSL) on a realistic benchmark where data exhibits considerable class imbalance and contains images from novel classes. Our benchmark consists of two fine-grained classification datasets obtained by sampling classes from the Aves and Fungi taxonomy. We find that recently proposed SSL methods provide significant benefits, and can effectively use out-of-class data to improve performance when deep networks are trained from scratch. Yet their performance pales in comparison to a transfer learning baseline, an alternative approach for learning from a few examples. Furthermore, in the transfer setting, while existing SSL methods provide improvements, the presence of out-of-class is often detrimental. In this setting, standard fine-tuning followed by distillation-based self-training is the most robust. Our work suggests that semi-supervised learning with experts on realistic datasets may require different strategies than those currently prevalent in the literature.
****
Residential Floor Plan Recognition and Reconstruction
Xiaolei Lv, Shengchu Zhao, Xinyang Yu, Binqiang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16717-16726
Recognition and reconstruction of residential floor plan drawings are important and challenging in design, decoration, and architectural remodeling fields. An automatic framework is provided that accurately recognizes the structure, type, and size of the room, and outputs vectorized 3D reconstruction results. Deep segmentation and detection neural networks are utilized to extract room structural information. Key points detection network and cluster analysis are utilized to calculate scales of rooms. The vectorization of room information is processed through an iterative optimization-based method. The system significantly increases accuracy and generalization ability, compared with existing methods. It outperforms other systems in floor plan segmentation and vectorization process, especially inclined wall detection.
****
Dynamic Domain Adaptation for Efficient Inference
Shuang Li, JinMing Zhang, Wenxuan Ma, Chi Harold Liu, Wei Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7832-7841
Domain adaptation (DA) enables knowledge transfer from a labeled source domain t

o an unlabeled target domain by reducing the cross-domain distribution discrepancy. Most prior DA approaches leverage complicated and powerful deep neural networks to improve the adaptation capacity and have shown remarkable success. However, they may have a lack of applicability to real-world situations such as real-time interaction, where low target inference latency is an essential requirement under limited computational budget. In this paper, we tackle the problem by proposing a dynamic domain adaptation (DDA) framework, which can simultaneously achieve efficient target inference in low-resource scenarios and inherit the favorable cross-domain generalization brought by DA. In contrast to static models, as a simple yet generic method, DDA can integrate various domain confusion constraints into any typical adaptive network, where multiple intermediate classifiers can be equipped to infer "easier" and "harder" target data dynamically. Moreover, we present two novel strategies to further boost the adaptation performance of multiple prediction exits: 1) a confidence score learning strategy to derive accurate target pseudo labels by fully exploring the prediction consistency of different classifiers; 2) a class-balanced self-training strategy to explicitly adapt multi-stage classifiers from source to target without losing prediction diversity. Extensive experiments on multiple benchmarks are conducted to verify that DDA can consistently improve the adaptation performance and accelerate target inference under domain shift and limited resources scenarios.
****
Regularization Strategy for Point Cloud via Rigidly Mixed Sample
Dogyoon Lee, Jaeha Lee, Junhyeop Lee, Hyeongmin Lee, Minhyeok Lee, Sungmin Woo, Sangyoun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15900-15909
Data augmentation is an effective regularization strategy to alleviate the overfitting, which is an inherent drawback of the deep neural networks. However, data augmentation is rarely considered for point cloud processing despite many studies proposing various augmentation methods for image data. Actually, regularization is essential for point clouds since lack of generality is more likely to occur in point cloud due to small datasets. This paper proposes a Rigid Subset Mix (RSMix), a novel data augmentation method for point clouds that generates a virtual mixed sample by replacing part of the sample with shape-preserved subsets from another sample. RSMix preserves structural information of the point cloud sample by extracting subsets from each sample without deformation using a neighboring function. The neighboring function was carefully designed considering unique properties of point cloud, unordered structure and non-grid. Experiments verified that RSMix successfully regularized the deep neural networks with remarkable improvement for shape classification. We also analyzed various combinations of data augmentations including RSMix with single and multi-view evaluations, based on abundant ablation studies.
****
StereoPIFu: Depth Aware Clothed Human Digitization via Stereo Vision
Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, Hujun Bao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 535-545
In this paper, we propose StereoPIFu, which integrates the geometric constraints of stereo vision with implicit function representation of PIFu, to recover the 3D shape of the clothed human from a pair of low-cost rectified images. First, we introduce the effective voxel-aligned features from a stereo vision-based network to enable depth-aware reconstruction. Moreover, the novel relative z-offset is employed to associate predicted high-fidelity human depth and occupancy inference, which helps restore fine-level surface details. Second, a network structure that fully utilizes the geometry information from the stereo images is designed to improve the human body reconstruction quality. Consequently, our StereoPIFu can naturally infer the human body's spatial location in camera space and maintain the correct relative position of different parts of the human body, which enables our method to capture human performance. Compared with previous works, our StereoPIFu significantly improves the robustness, completeness, and accuracy of the clothed human reconstruction, which is demonstrated by extensive experiment

al results.
****
Unsupervised Multi-Source Domain Adaptation Without Access to Source Data
Sk Miraj Ahmed, Dripta S. Raychaudhuri, Sujoy Paul, Samet Oymak, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10103-10112
Unsupervised Domain Adaptation (UDA) aims to learn a predictor model for an unlabeled dataset by transferring knowledge from a labeled source data, which has been trained on similar tasks. However, most of these conventional UDA approaches have a strong assumption of having access to the source data during training, which may not be very practical due to privacy, security and storage concerns. A recent line of work addressed this problem and proposed an algorithm that transfers knowledge to the unlabeled target domain only from a single learned source model without requiring access to the source data. However, for adaptation purpose, if there are multiple trained source models available to choose from, this method has to go through adapting each and every model individually, to check for the best source. Thus, we ask the question: can we find the optimal combination of source models, with no source data and without target labels, whose performance is no worse than the single best source? To answer this, we propose a novel and efficient algorithm which automatically combines the source models with suitable weights in such a way that it performs at least as good as the best source model. We provide intuitive theoretical insights to justify our claim. Moreover, extensive experiments are conducted on several benchmark datasets to show the effectiveness of our algorithm, where in most cases, our method not only reaches best source accuracy but also outperform it.
****
On Semantic Similarity in Video Retrieval
Michael Wray, Hazel Doughty, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3650-3660
Current video retrieval efforts all found their evaluation on an instance-based assumption, that only a single caption is relevant to a query video and vice versa. We demonstrate that this assumption results in performance comparisons often not indicative of models' retrieval capabilities. We propose a move to semantic similarity video retrieval, where (i) multiple videos/captions can be deemed equally relevant, and their relative ranking does not affect a method's reported performance and (ii) retrieved videos/captions are ranked by their similarity to a query. We propose several proxies to estimate semantic similarities in large-scale retrieval datasets, without additional annotations. Our analysis is performed on three commonly used video retrieval datasets (MSR-VTT, YouCook2 and EPIC-KITCHENS).
****
Few-Shot Open-Set Recognition by Transformation Consistency
Minki Jeong, Seokeon Choi, Changick Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12566-12575
In this paper, we attack a few-shot open-set recognition (FSOSR) problem, which is a combination of few-shot learning (FSL) and open-set recognition (OSR). It aims to quickly adapt a model to a given small set of labeled samples while rejecting unseen class samples. Since OSR requires rich data and FSL considers closed-set classification, existing OSR and FSL methods show poor performances in solving FSOSR problems. The previous FSOSR method utilizes pseudo-unseen class samples, which are collected from the other dataset or synthesized samples to model unseen class representations. However, this approach is heavily dependent on the composition of the pseudo samples. In this paper, we propose a novel unknown class sample detector, named SnaTCHer, that does not require pseudo-unseen samples. Based on the transformation consistency, our method measures the difference between the transformed prototypes and a modified prototype set. The modified set is composed by replacing a query feature and its predicted class prototype. SnaTCHer rejects samples with large differences to the transformed prototypes. Our method alters the unseen class distribution estimation problem to a relative feature transformation problem, independent of pseudo-unseen class samples. We invest

igate our SnaTCHer with various prototype transformation methods and observe that our method consistently improves unseen class sample detection performance without closed-set classification reduction.
****
Uncertainty-Guided Model Generalization to Unseen Domains
Fengchun Qiao, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6790-6800
We study a worst-case scenario in generalization: Out-of-domain generalization from a single source. The goal is to learn a robust model from a single source and expect it to generalize over many unknown distributions. This challenging problem has been seldom investigated while existing solutions suffer from various limitations. In this paper, we propose a new solution. The key idea is to augment the source capacity in both input and label spaces, while the augmentation is guided by uncertainty assessment. To the best of our knowledge, this is the first work to (1) access the generalization uncertainty from a single source and (2) leverage it to guide both input and label augmentation for robust generalization. The model training and deployment are effectively organized in a Bayesian meta-learning framework. We conduct extensive comparisons and ablation study to validate our approach. The results prove our superior performance in a wide scope of tasks including image classification, semantic segmentation, text classification, and speech recognition.
****
Debiased Subjective Assessment of Real-World Image Enhancement
Peibei Cao, Zhangyang Wang, Kede Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 711-721
In real-world image enhancement, it is often challenging (if not impossible) to acquire ground-truth data, preventing the adoption of distance metrics for objective quality assessment. As a result, one often resorts to subjective quality assessment, the most straightforward and reliable means of evaluating image enhancement. Conventional subjective testing requires manually pre-selecting a small set of visual examples, which may suffer from three sources of biases: 1) sampling bias due to the extremely sparse distribution of the selected samples in the image space; 2) algorithmic bias due to potential overfitting the selected samples; 3) subjective bias due to further potential cherry-picking test results. This eventually makes the field of real-world image enhancement more of an art than a science. Here we take steps towards debiasing conventional subjective assessment by automatically sampling a set of adaptive and diverse images for subsequent testing. This is achieved by casting sample selection into a joint maximization of the discrepancy between the enhancers and the diversity among the selected input images. Careful visual inspection on the resulting enhanced images provides a debiased ranking of the enhancement algorithms. We demonstrate our subjective assessment method using three popular and practically demanding image enhancement tasks: dehazing, super-resolution, and low-light enhancement.
****
Landmark Regularization: Ranking Guided Super-Net Training in Neural Architecture Search
Kaicheng Yu, Rene Ranftl, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13723-13732
Weight sharing has become a de facto standard in neural architecture search because it enables the search to be done on commodity hardware. However, recent works have empirically shown a ranking disorder between the performance of stand-alone architectures and that of the corresponding shared-weight networks. This violates the main assumption of weight-sharing NAS algorithms, thus limiting their effectiveness. We tackle this issue by proposing a regularization term that aims to maximize the correlation between the performance rankings of the shared-weight network and that of the standalone architectures using a small set of landmark architectures. We incorporate our regularization term into three different NAS algorithms and show that it consistently improves performance across algorithms, search-spaces, and tasks.
****

Noise-Resistant Deep Metric Learning With Ranking-Based Instance Selection
Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, Chunyan Miao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6811-6820
The existence of noisy labels in real-world data negatively impacts the performance of deep learning models. Although much research effort has been devoted to improving robustness to noisy labels in classification tasks, the problem of noisy labels in deep metric learning (DML) remains open. In this paper, we propose a noise-resistant training technique for DML, which we name Probabilistic Ranking-based Instance Selection with Memory (PRISM). PRISM identifies noisy data in a minibatch using average similarity against image features extracted by several previous versions of the neural network. These features are stored in and retrieved from a memory bank. To alleviate the high computational cost brought by the memory bank, we introduce an acceleration method that replaces individual data points with the class centers. In extensive comparisons with 12 existing approaches under both synthetic and real-world label noise, PRISM demonstrates superior performance of up to 6.06% in Precision@1.
****
Neural Reprojection Error: Merging Feature Learning and Camera Pose Estimation
Hugo Germain, Vincent Lepetit, Guillaume Bourmaud; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 414-423
Absolute camera pose estimation is usually addressed by sequentially solving two distinct subproblems: First a feature matching problem that seeks to establish putative 2D-3D correspondences, and then a Perspective-n-Point problem that minimizes, w.r.t. the camera pose, the sum of so-called Reprojection Errors (RE). We argue that generating putative 2D-3D correspondences 1) leads to an important loss of information that needs to be compensated as far as possible, within RE, through the choice of a robust loss and the tuning of its hyperparameters and 2) may lead to an RE that conveys erroneous data to the pose estimator. In this paper, we introduce the Neural Reprojection Error (NRE) as a substitute for RE. NRE allows to rethink the camera pose estimation problem by merging it with the feature learning problem, hence leveraging richer information than 2D-3D correspondences and eliminating the need for choosing a robust loss and its hyperparameters. Thus NRE can be used as training loss to learn image descriptors tailored for pose estimation. We also propose a coarse-to-fine optimization method able to very efficiently minimize a sum of NRE terms w.r.t. the camera pose. We experimentally demonstrate that NRE is a good substitute for RE as it significantly improves both the robustness and the accuracy of the camera pose estimate while being computationally and memory highly efficient. From a broader point of view, we believe this new way of merging deep learning and 3D geometry may be useful in other computer vision applications.
****
Cross Modal Focal Loss for RGBD Face Anti-Spoofing
Anjith George, Sebastien Marcel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7882-7891
Automatic methods for detecting presentation attacks are essential to ensure the reliable use of facial recognition technology. Most of the methods available in the literature for presentation attack detection (PAD) fails in generalizing to unseen attacks. In recent years, multi-channel methods have been proposed to improve the robustness of PAD systems. Often, only a limited amount of data is available for additional channels, which limits the effectiveness of these methods. In this work, we present a new framework for PAD that uses RGB and depth channels together with a novel loss function. The new architecture uses complementary information from the two modalities while reducing the impact of overfitting. Essentially, a cross-modal focal loss function is proposed to modulate the loss contribution of each channel as a function of the confidence of individual channels. Extensive evaluations in two publicly available datasets demonstrate the effectiveness of the proposed approach.
****
StickyPillars: Robust and Efficient Feature Matching on Point Clouds Using Graph

Neural Networks
Kai Fischer, Martin Simon, Florian Olsner, Stefan Milz, Horst-Michael Gross, Patrick Mader; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 313-323
Robust point cloud registration in real-time is an important prerequisite for many mapping and localization algorithms. Traditional methods like ICP tend to fail without good initialization, insufficient overlap or in the presence of dynamic objects. Modern deep learning based registration approaches present much better results, but suffer from a heavy runtime. We overcome these drawbacks by introducing StickyPillars, a fast, accurate and extremely robust deep middle-end 3D feature matching method on point clouds. It uses graph neural networks and performs context aggregation on sparse 3D key-points with the aid of transformer based multi-head self and cross-attention. The network output is used as the cost for an optimal transport problem whose solution yields the final matching probabilities. The system does not rely on hand crafted feature descriptors or heuristic matching strategies. We present state-of-art art accuracy results on the registration problem demonstrated on the KITTI dataset while being four times faster then leading deep methods. Furthermore, we integrate our matching system into a LiDAR odometry pipeline yielding most accurate results on the KITTI odometry dataset. Finally, we demonstrate robustness on KITTI odometry. Our method remains stable in accuracy where state-of-the-art procedures fail on frame drops and higher speeds.
****
HoHoNet: 360 Indoor Holistic Understanding With Latent Horizontal Features
Cheng Sun, Min Sun, Hwann-Tzong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2573-2582
We present HoHoNet, a versatile and efficient framework for holistic understanding of an indoor 360-degree panorama using a Latent Horizontal Feature (LHFeat). The compact LHFeat flattens the features along the vertical direction and has shown success in modeling per-column modality for room layout reconstruction. HoHoNet advances in two important aspects. First, the deep architecture is redesigned to run faster with improved accuracy. Second, we propose a novel horizon-to-dense module, which relaxes the per-column output shape constraint, allowing per-pixel dense prediction from LHFeat. HoHoNet is fast: It runs at 52 FPS and 110 FPS with ResNet-50 and ResNet-34 backbones respectively, for modeling dense modalities from a high-resolution 512x1024 panorama. HoHoNet is also accurate. On the tasks of layout estimation and semantic segmentation, HoHoNet achieves results on par with current state-of-the-art. On dense depth estimation, HoHoNet outperforms all the prior arts by a large margin.
****
Online Learning of a Probabilistic and Adaptive Scene Representation
Zike Yan, Xin Wang, Hongbin Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13111-13121
Constructing and maintaining a consistent scene model on-the-fly is the core task for online spatial perception, interpretation, and action. In this paper, we represent the scene with a Bayesian nonparametric mixture model, seamlessly describing per-point occupancy status with a continuous probability density function. Instead of following the conventional data fusion paradigm, we address the problem of online learning the process how sequential point cloud data are generated from the scene geometry. An incremental and parallel inference is performed to update the parameter space in real-time. We experimentally show that the proposed representation achieves state-of-the-art accuracy with promising efficiency. The consistent probabilistic formulation assures a generative model that is adaptive to different sensor characteristics, and the model complexity can be dynamically adjusted on-the-fly according to different data scales.
****
Domain Adaptation With Auxiliary Target Domain-Oriented Classifier
Jian Liang, Dapeng Hu, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16632-16642
Domain adaptation (DA) aims to transfer knowledge from a label-rich but heteroge

neous domain to a label-scare domain, which alleviates the labeling efforts and attracts considerable attention. Different from previous methods focusing on learning domain-invariant feature representations, some recent methods present generic semi-supervised learning (SSL) techniques and directly apply them to DA tasks, even achieving competitive performance. One of the most popular SSL techniques is pseudo-labeling that assigns pseudo labels for each unlabeled data via the classifier trained by labeled data. However, it ignores the distribution shift in DA problems and is inevitably biased to source data. To address this issue, we propose a new pseudo-labeling framework called Auxiliary Target Domain-Oriented Classifier (ATDOC). ATDOC alleviates the classifier bias by introducing an auxiliary classifier for target data only, to improve the quality of pseudo labels. Specifically, we employ the memory mechanism and develop two types of non-parametric classifiers, i.e. the nearest centroid classifier and neighborhood aggregation, without introducing any additional network parameters. Despite its simplicity in a pseudo classification objective, ATDOC with neighborhood aggregation significantly outperforms domain alignment techniques and prior SSL techniques on a large variety of DA benchmarks and even scare-labeled SSL tasks.
****
Learning To Recover 3D Scene Shape From a Single Image
Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 204-213
Despite significant progress in monocular depth estimation in the wild, recent state-of-the-art methods cannot be used to recover accurate 3D scene shape due to an unknown depth shift induced by shift-invariant reconstruction losses used in mixed-data depth prediction training, and possible unknown camera focal length. We investigate this problem in detail and propose a two-stage framework that first predicts depth up to an unknown scale and shift from a single monocular image, and then use 3D point cloud encoders to predict the missing depth shift and focal length that allow us to recover a realistic 3D scene shape. In addition, we propose an image-level normalized regression loss and a normal-based geometry loss to enhance depth prediction models trained on mixed datasets. We test our depth model on nine unseen datasets and achieve state-of-the-art performance on zero-shot dataset generalization. Code is available at:https://git.io/Depth.
****
Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes
Zhengqi Li, Simon Niklaus, Noah Snavely, Oliver Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6498-6508
We present a method to perform novel view and time synthesis of dynamic scenes, requiring only a monocular video with known camera poses as input. To do this, we introduce Neural Scene Flow Fields, a new representation that models the dynamic scene as a time-variant continuous function of appearance, geometry, and 3D scene motion. Our representation is optimized through a neural network to fit the observed input views. We show that our representation can be used for complex dynamic scenes, including thin structures, view-dependent effects, and natural degrees of motion. We conduct a number of experiments that demonstrate our approach significantly outperforms recent monocular view synthesis methods, and show qualitative results of space-time view synthesis on a variety of real-world videos.
****
FS-Net: Fast Shape-Based Network for Category-Level 6D Object Pose Estimation With Decoupled Rotation Mechanism
Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, Ales Leonardis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1581-1590
In this paper, we focus on category-level 6D pose and size estimation from a monocular RGB-D image. Previous methods suffer from inefficient category-level pose feature extraction, which leads to low accuracy and inference speed. To tackle this problem, we propose a fast shape-based network (FS-Net) with efficient cate

gory-level feature extraction for 6D pose estimation. First, we design an orient ation aware autoencoder with 3D graph convolution for latent feature extraction. Thanks to the shift and scale-invariance properties of 3D graph convolution, th e learned latent feature is insensitive to point shift and object size. Then, to efficiently decode category-level rotation information from the latent feature, we propose a novel decoupled rotation mechanism that employs two decoders to co mplementarily access the rotation information. For translation and size, we esti mate them by two residuals: the difference between the mean of object points and ground truth translation, and the difference between the mean size of the categ ory and ground truth size, respectively. Finally, to increase the generalization ability of the FS-Net, we propose an online box-cage based 3D deformation mecha nism to augment the training data. Extensive experiments on two benchmark datase ts show that the proposed method achieves state-of-the-art performance in both c ategory- and instance-level 6D object pose estimation. Especially in category-le vel pose estimation, without extra synthetic data, our method outperforms existi ng methods by 6.3% on the NOCS-REAL dataset.
****
Unsupervised Human Pose Estimation Through Transforming Shape Templates
Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi , Bernhard Kainz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2484-2494
Human pose estimation is a major computer vision problem with applications rangi ng from augmented reality and video capture to surveillance and movement trackin g. In the medical context, the latter may be an important biomarker for neurolog ical impairments in infants. Whilst many methods exist, their application has be en limited by the need for well annotated large datasets and the inability to ge neralize to humans of different shapes and body compositions, e.g. children and infants. In this paper we present a novel method for learning pose estimators fo r human adults and infants in an unsupervised fashion. We approach this as a lea rnable template matching problem facilitated by deep feature extractors. Human-i nterpretable landmarks are estimated by transforming a template consisting of pr edefined body parts that are characterized by 2D Gaussian distributions. Enforci ng a connectivity prior guides our model to meaningful human shape representatio ns. We demonstrate the effectiveness of our approach on two different datasets i ncluding adults and infants.
****
Improving OCR-Based Image Captioning by Incorporating Geometrical Relationship
Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, Jiebo Luo; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1306-1315
OCR-based image captioning aims to automatically describe images based on all th e visual entities (both visual objects and scene text) in images. Compared with conventional image captioning, the reasoning of scene text is required for OCR-b ased image captioning since the generated descriptions often contain multiple OC R tokens. Existing methods attempt to achieve this goal via encoding the OCR tok ens with rich visual and semantic representations. However, strong correlations between OCR tokens may not be established with such limited representations. In this paper, we propose to enhance the connections between OCR tokens from the vi ewpoint of exploiting the geometrical relationship. We comprehensively consider the height, width, distance, IoU and orientation relations between the OCR token s for constructing the geometrical relationship. To integrate the learned relati on as well as the visual and semantic representations into a unified framework, a Long Short-Term Memory plus Relation-aware pointer network (LSTM-R) architectu re is presented in this paper. Under the guidance of the geometrical relationshi p between OCR tokens, our LSTM-R capitalizes on a newly-devised relation-aware p ointer network to select OCR tokens from the scene text for OCR-based image capt ioning. Extensive experiments demonstrate the effectiveness of our LSTM-R. More remarkably, LSTM-R achieves state-of-the-art performance on TextCaps, with the C IDEr-D score being increased from 98.0% to 109.3%.
****

Cross-Iteration Batch Normalization
Zhuliang Yao, Yue Cao, Shuxin Zheng, Gao Huang, Stephen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12331-12340
A well-known issue of Batch Normalization is its significantly reduced effectiveness in the case of small mini-batch sizes. When a mini-batch contains few examples, the statistics upon which the normalization is defined cannot be reliably estimated from it during a training iteration. To address this problem, we present Cross-Iteration Batch Normalization (CBN), in which examples from multiple recent iterations are jointly utilized to enhance estimation quality. A challenge of computing statistics over multiple iterations is that the network activations from different iterations are not comparable to each other due to changes in network weights. We thus compensate for the network weight changes via a proposed technique based on Taylor polynomials, so that the statistics can be accurately estimated and batch normalization can be effectively applied. On object detection and image classification with small mini-batch sizes, CBN is found to outperform the original batch normalization and a direct calculation of statistics over previous iterations without the proposed compensation technique.
****
Multimodal Contrastive Training for Visual Representation Learning
Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, Baldo Faieta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6995-7004
We develop an approach to learning visual representations that embraces multimodal data, driven by a combination of intra- and inter-modal similarity preservation objectives. Unlike existing visual pre-training methods, which solve a proxy prediction task in a single domain, our method exploits intrinsic data properties within each modality and semantic information from cross-modal correlation simultaneously, hence improving the quality of learned visual representations. By including multimodal training in a unified framework with different types of contrastive losses, our method can learn more powerful and generic visual features. We first train our model on COCO and evaluate the learned visual representations on various downstream tasks including image classification, object detection, and instance segmentation. For example, the visual representations pre-trained on COCO by our method achieve state-of-the-art top-1 validation accuracy of 55.3% on ImageNet classification, under the common transfer protocol. We also evaluate our method on the large-scale Stock images dataset and show its effectiveness on multi-label image tagging, and cross-modal retrieval tasks.
****
3D Shape Generation With Grid-Based Implicit Functions
Moritz Ibing, Isaak Lim, Leif Kobbelt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13559-13568
Previous approaches to generate shapes in a 3D setting train a GAN on the latent space of an autoencoder (AE). Even though this produces convincing results, it has two major shortcomings. As the GAN is limited to reproduce the dataset the AE was trained on, we cannot reuse a trained AE for novel data. Furthermore, it is difficult to add spatial supervision into the generation process, as the AE only gives us a global representation. To remedy these issues, we propose to train the GAN on grids (i.e. each cell covers a part of a shape). In this representation each cell is equipped with a latent vector provided by an AE. This localized representation enables more expressiveness (since the cell-based latent vectors can be combined in novel ways) as well as spatial control of the generation process (e.g. via bounding boxes). Our method outperforms the current state of the art on all established evaluation measures, proposed for quantitatively evaluating the generative capabilities of GANs. We show limitations of these measures and propose the adaptation of a robust criterion from statistical analysis as an alternative.
****
Tangent Space Backpropagation for 3D Transformation Groups
Zachary Teed, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer Visio

n and Pattern Recognition (CVPR), 2021, pp. 10338-10347
We address the problem of performing backpropagation for computation graphs involving 3D transformation groups SO(3), SE(3), and Sim(3). 3D transformation groups are widely used in 3D vision and robotics, but they do not form vector spaces and instead lie on smooth manifolds. The standard backpropagation approach, which embeds 3D transformations in Euclidean spaces, suffers from numerical difficulties. We introduce a new library, which exploits the group structure of 3D transformations and performs backpropagation in the tangent spaces of manifolds. We show that our approach is numerically more stable, easier to implement, and beneficial to a diverse set of tasks. Our plug-and-play PyTorch library is available at https://github.com/princeton-vl/lietorch.
****
FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation
Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14050-14059
Image caption evaluation is a crucial task, which involves the semantic perception and matching of image and text. Good evaluation metrics aim to be fair, comprehensive, and consistent with human judge intentions. When humans evaluate a caption, they usually consider multiple aspects, such as whether it is related to the target image without distortion, how much image gist it conveys, as well as how fluent and beautiful the language and wording is. The above three different evaluation orientations can be summarized as fidelity, adequacy, and fluency. The former two rely on the image content, while fluency is purely related to linguistics and more subjective. Inspired by human judges, we propose a learning-based metric named FAIEr to ensure evaluating the fidelity and adequacy of the captions. Since image captioning involves two different modalities, we employ the scene graph as a bridge between them to represent both images and captions. FAIEr mainly regards the visual scene graph as the criterion to measure the fidelity. Then for evaluating the adequacy of the candidate caption, it highlights the image gist on the visual scene graph under the guidance of the reference captions. Comprehensive experimental results show that FAIEr has high consistency with human judgment as well as high stability, low reference dependency, and the capability of reference-free evaluation.
****
HLA-Face: Joint High-Low Adaptation for Low Light Face Detection
Wenjing Wang, Wenhan Yang, Jiaying Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16195-16204
Face detection in low light scenarios is challenging but vital to many practical applications, e.g., surveillance video, autonomous driving at night. Most existing face detectors heavily rely on extensive annotations, while collecting data is time-consuming and laborious. To reduce the burden of building new datasets for low light conditions, we make full use of existing normal light data and explore how to adapt face detectors from normal light to low light. The challenge of this task is that the gap between normal and low light is too huge and complex for both pixel-level and object-level. Therefore, most existing low-light enhancement and adaptation methods do not achieve desirable performance. To address the issue, we propose a joint High-Low Adaptation (HLA) framework. Through a bidirectional low-level adaptation and multi-task high-level adaptation scheme, our HLA-Face outperforms state-of-the-art methods even without using dark face labels for training. Our project is publicly available at: https://daooshee.github.io/HLA-Face-Website/
****
Hierarchical Video Prediction Using Relational Layouts for Human-Object Interactions
Navaneeth Bodla, Gaurav Shrivastava, Rama Chellappa, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12146-12155
Learning to model and predict how humans interact with objects while performing an action is challenging, and most of the existing video prediction models are i

neffective in modeling complicated human-object interactions. Our work builds on hierarchical video prediction models, which disentangle the video generation process into two stages: predicting a high-level representation, such as pose sequence, and then learning a pose-to-pixels translation model for pixel generation. An action sequence for a human-object interaction task is typically very complicated, involving the evolution of pose, person's appearance, object locations, and object appearances over time. To this end, we propose a Hierarchical Video Prediction model using Relational Layouts. In the first stage, we learn to predict a sequence of layouts. A layout is a high-level representation of the video containing both pose and objects' information for every frame. The layout sequence is learned by modeling the relationships between the pose and objects using relational reasoning and recurrent neural networks. The layout sequence acts as a strong structure prior to the second stage that learns to map the layouts into pixel space. Experimental evaluation of our method on two datasets, UMD-HOI and Bimanual, shows significant improvements in standard video evaluation metrics such as LPIPS, PSNR, and SSIM. We also perform a detailed qualitative analysis of our model to demonstrate various generalizations.
****

From Rain Generation to Rain Removal

Hong Wang, Zongsheng Yue, Qi Xie, Qian Zhao, Yefeng Zheng, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14791-14801

For the single image rain removal (SIRR) task, the performance of deep learning (DL)-based methods is mainly affected by the designed deraining models and training datasets. Most of current state-of-the-art focus on constructing powerful deep models to obtain better deraining results. In this paper, to further improve the deraining performance, we novelly attempt to handle the SIRR task from the perspective of training datasets by exploring a more efficient way to synthesize rainy images. Specifically, we build a full Bayesian generative model for rainy image where the rain layer is parameterized as a generator with the input as some latent variables representing the physical structural rain factors, e.g., direction, scale, and thickness. To solve this model, we employ the variational inference framework to approximate the expected statistical distribution of rainy image in a data-driven manner. With the learned generator, we can automatically and sufficiently generate diverse and non-repetitive training pairs so as to efficiently enrich and augment the existing benchmark datasets. User study qualitatively and quantitatively evaluates the realism of generated rainy images. Comprehensive experiments substantiate that the proposed model can faithfully extract the complex rain distribution that not only helps significantly improve the deraining performance of current deep single image derainers, but also largely loosens the requirement of large training sample pre-collection for the SIRR task. Code is available in https://github.com/hongwang01/VRGNet.
****

Few-Shot Classification With Feature Map Reconstruction Networks

Davis Wertheimer, Luming Tang, Bharath Hariharan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8012-8021

In this paper we reformulate few-shot classification as a reconstruction problem in latent space. The ability of the network to reconstruct a query feature map from support features of a given class predicts membership of the query in that class. We introduce a novel mechanism for few-shot classification by regressing directly from support features to query features in closed form, without introducing any new modules or large-scale learnable parameters. The resulting Feature Map Reconstruction Networks are both more performant and computationally efficient than previous approaches. We demonstrate consistent and substantial accuracy gains on four fine-grained benchmarks with varying neural architectures. Our model is also competitive on the non-fine-grained mini-ImageNet and tiered-ImageNet benchmarks with minimal bells and whistles.
****

Object Classification From Randomized EEG Trials

Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, Jeffrey Mark Siskind; Proceedi

ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3845-3854
New results suggest strong limits to the feasibility of object classification from human brain activity evoked by image stimuli, as measured through EEG. Considerable prior work suffers from a confound between the stimulus class and the time since the start of the experiment. A prior attempt to avoid this confound using randomized trials was unable to achieve results above chance in a statistically significant fashion when the data sets were of the same size as the original experiments. Here, we attempt object classification from EEG using an array of methods that are representative of the state-of-the-art, with a far larger (20x) dataset of randomized EEG trials, 1,000 stimulus presentations of each of forty classes, all from a single subject. To our knowledge, this is the largest such EEG data-collection effort from a single subject and is at the bounds of feasibility. We obtain classification accuracy that is marginally above chance and above chance in a statistically significant fashion, and further assess how accuracy depends on the classifier used, the amount of training data used, and the number of classes. Reaching the limits of data collection with only marginally above-chance performance suggests that the prevailing literature substantially exaggerates the feasibility of object classification from EEG.
****
Learning Monocular 3D Reconstruction of Articulated Categories From Motion
Filippos Kokkinos, Iasonas Kokkinos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1737-1746
Monocular 3D reconstruction of articulated object categories is challenging due to the lack of training data and the inherent ill-posedness of the problem. In this work we use video self-supervision, forcing the consistency of consecutive 3D reconstructions by a motion-based cycle loss. This largely improves both optimization-based and learning-based 3D mesh reconstruction. We further introduce an interpretable model of 3D template deformations that controls a 3D surface through the displacement of a small number of local, learnable handles. We formulate this operation as a structured layer relying on mesh-laplacian regularization and show that it can be trained in an end-to-end manner. We finally introduce a per-sample numerical optimisation approach that jointly optimises over mesh displacements and cameras within a video, boosting accuracy both for training and also as test time post-processing. While relying exclusively on a small set of videos collected per category for supervision, we obtain state-of-the-art reconstructions with diverse shapes, viewpoints and textures for multiple articulated object categories.
****
De-Rendering the World's Revolutionary Artefacts
Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6338-6347
Recent works have shown exciting results in unsupervised image de-rendering--learning to decompose 3D shape, appearance, and lighting from single-image collections without explicit supervision. However, many of these assume simplistic material and lighting models. We propose a method, termed RADAR, that can recover environment illumination and surface materials from real single-image collections, relying neither on explicit 3D supervision, nor on multi-view or multi-light images. Specifically, we focus on rotationally symmetric artefacts that exhibit challenging surface properties including specular reflections, such as vases. We introduce a novel self-supervised albedo discriminator, which allows the model to recover plausible albedo without requiring any ground-truth during training. In conjunction with a shape reconstruction module exploiting rotational symmetry, we present an end-to-end learning framework that is able to de-render the world's revolutionary artefacts. We conduct experiments on a real vase dataset and demonstrate compelling decomposition results, allowing for applications including free-viewpoint rendering and relighting. More results and code at: https://sorderender.github.io/.
****

Progressively Complementary Network for Fisheye Image Rectification Using Appearance Flow
Shangrong Yang, Chunyu Lin, Kang Liao, Chunjie Zhang, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6348-6357
Distortion rectification is often required for fisheye images. The generation-based method is one mainstream solution due to its label-free property, but its naive skip-connection and overburdened decoder will cause blur and incomplete correction. First, the skip-connection directly transfers the image features, which may introduce distortion and cause incomplete correction. Second, the decoder is overburdened during simultaneously reconstructing the content and structure of the image, resulting in vague performance. To solve these two problems, in this paper, we focus on the interpretable correction mechanism of the distortion rectification network and propose a feature-level correction scheme. We embed a correction layer in skip-connection and leverage the appearance flows in different layers to pre-correct the image features. Consequently, the decoder can easily reconstruct a plausible result with the remaining distortion-less information. In addition, we propose a parallel complementary structure. It effectively reduces the burden of the decoder by separating content reconstruction and structure correction. Subjective and objective experiment results on different datasets demonstrate the superiority of our method.
****
DECOR-GAN: 3D Shape Detailization by Conditional Refinement
Zhiqin Chen, Vladimir G. Kim, Matthew Fisher, Noam Aigerman, Hao Zhang, Siddhartha Chaudhuri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15740-15749
We introduce a deep generative network for 3D shape detailization, akin to stylization with the style being geometric details. We address the challenge of creating large varieties of high-resolution and detailed 3D geometry from a small set of exemplars by treating the problem as that of geometric detail transfer. Given a low-resolution coarse voxel shape, our network refines it, via voxel upsampling, into a higher-resolution shape enriched with geometric details. The output shape preserves the overall structure (or content) of the input, while its detail generation is conditioned on an input "style code" corresponding to a detailed exemplar. Our 3D detailization via conditional refinement is realized by a generative adversarial network, coined DECOR-GAN. The network utilizes a 3D CNN generator for upsampling coarse voxels and a 3D PatchGAN discriminator to enforce local patches of the generated model to be similar to those in the training detailed shapes. During testing, a style code is fed into the generator to condition the refinement. We demonstrate that our method can refine a coarse shape into a variety of detailed shapes with different styles. The generated results are evaluated in terms of content preservation, plausibility, and diversity. Comprehensive ablation studies are conducted to validate our network designs. Code is available at https://github.com/czq142857/DECOR-GAN.
****
Model-Aware Gesture-to-Gesture Translation
Hezhen Hu, Weilun Wang, Wengang Zhou, Weichao Zhao, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16428-16437
Hand gesture-to-gesture translation is a significant and interesting problem, which serves as a key role in many applications, such as sign language production. This task involves fine-grained structure understanding of the mapping between the source and target gestures. Current works follow a data-driven paradigm based on sparse 2D joint representation. However, given the insufficient representation capability of 2D joints, this paradigm easily leads to blurry generation results with incorrect structure. In this paper, we propose a novel model-aware gesture-to-gesture translation framework, which introduces hand prior with hand meshes as the intermediate representation. To take full advantage of the structured hand model, we first build a dense topology map aligning the image plane with the encoded embedding of the visible hand mesh. Then, a transformation flow is ca

lculated based on the correspondence of the source and target topology map. During the generation stage, we inject the topology information into generation streams by modulating the activations in a spatially-adaptive manner. Further, we incorporate the source local characteristic to enhance the translated gesture image according to the transformation flow. Extensive experiments on two benchmark datasets have demonstrated that our method achieves new state-of-the-art performance.
****
Spatio-temporal Contrastive Domain Adaptation for Action Recognition
Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, Hua Chai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9787-9795
Unsupervised domain adaptation (UDA) for human action recognition is a practical and challenging problem. Compared with image-based UDA, video-based UDA is comprehensive to bridge the domain shift on both spatial representation and temporal dynamics. Most previous works focus on short-term modeling and alignment with frame-level or clip-level features, which is not discriminative sufficiently for video-based UDA tasks. To address these problems, in this paper we propose to establish the cross-modal domain alignment via self-supervised contrastive framework, i.e., spatio-temporal contrastive domain adaptation (STCDA), to learn the joint clip-level and video-level representation alignment. Since the effective representation is modeled from unlabeled data by self-supervised learning (SSL), spatio-temporal contrastive learning (STCL) is proposed to explore the useful long-term feature representation for classification, using self-supervision setting trained from the contrastive clip/video pairs with positive or negative properties. Besides, we involve a novel domain metric scheme, i.e., video-based contrastive alignment (VCA), to optimize the category-aware video-level alignment and generalization between source and target. The proposed STCDA achieves stat-of-the-art results on several UDA benchmarks for action recognition.
****
Exploiting Semantic Embedding and Visual Feature for Facial Action Unit Detection
Huiyuan Yang, Lijun Yin, Yi Zhou, Jiuxiang Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10482-10491
Recent study on detecting facial action units (AU) has utilized auxiliary information (i.e., facial landmarks, relationship among AUs and expressions, web facial images, etc.), in order to improve the AU detection performance. As of now, no semantic information of AUs has yet been explored for such a task. As a matter of fact, AU semantic descriptions provide much more information than the binary AU labels alone, thus we propose to exploit the Semantic Embedding and Visual feature (SEV-Net) for AU detection. More specifically, AU semantic embeddings are obtained through both Intra-AU and Inter-AU attention modules, where the Intra-AU attention module captures the relation among words within each sentence that describes individual AU, and the Inter-AU attention module focuses on the relation among those sentences. The learned AU semantic embeddings are then used as guidance for the generation of attention maps through a cross-modality attention network. The generated cross-modality attention maps are further used as weights for the aggregated feature. Our proposed method is unique in that the semantic features are exploited as the first of this kind. The approach has been evaluated on three public AU-coded facial expression databases, and has achieved a superior performance than the state-of-the-art peer methods.
****
Categorical Depth Distribution Network for Monocular 3D Object Detection
Cody Reading, Ali Harakeh, Julia Chae, Steven L. Waslander; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8555-8564
Monocular 3D object detection is a key problem for autonomous vehicles, as it provides a solution with simple configuration compared to typical multi-sensor systems. The main challenge in monocular 3D detection lies in accurately predicting object depth, which must be inferred from object and scene cues due to the lack

of direct range measurement. Many methods attempt to directly estimate depth to assist in 3D detection, but show limited performance as a result of depth inaccuracy. Our proposed solution, Categorical Depth Distribution Network (CaDDN), uses a predicted categorical depth distribution for each pixel to project rich contextual feature information to the appropriate depth interval in 3D space. We then use the computationally efficient bird's-eye-view projection and single-stage detector to produce the final output bounding boxes. We design CaDDN as a fully differentiable end-to-end approach for joint depth estimation and object detection. We validate our approach on the KITTI 3D object detection benchmark, where we rank 1st among published monocular methods. We also provide the first monocular 3D detection results on the newly released Waymo Open Dataset. We provide a code release for CaDDN which will be made publicly available.
****
Learning From the Master: Distilling Cross-Modal Advanced Knowledge for Lip Reading
Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13325-13333
Lip reading aims to predict the spoken sentences from silent lip videos. Due to the fact that such a vision task usually performs worse than its counterpart speech recognition, one potential scheme is to distill knowledge from a teacher pretrained by audio signals. However, the latent domain gap between the cross-modal data could lead to an learning ambiguity and thus limits the performance of lip reading. In this paper, we propose a novel collaborative framework for lip reading, and two aspects of issues are considered: 1) the teacher should understand bi-modal knowledge to possibly bridge the inherent cross-modal gap; 2) the teacher should adjust teaching contents adaptively with the evolution of the student. To these ends, we introduce a trainable "master" network which ingests both audio signals and silent lip videos instead of a pretrained teacher. The master produces logits from three modalities of features: audio modality, video modality, and their combination. To further provide an interactive strategy to fuse these knowledge organically, we regularize the master with the task-specific feedback from the student, in which the requirement of the student is implicitly embedded. Meanwhile we involve a couple of "tutor" networks into our system as guidance for emphasizing the fruitful knowledge flexibly. In addition, we incorporate a curriculum learning design to ensure a better convergence. Extensive experiments demonstrate that the proposed network outperforms the state-of-the-art methods on several benchmarks, including in both word-level and sentence-level scenarios.
****
Spatially-Varying Outdoor Lighting Estimation From Intrinsics
Yongjie Zhu, Yinda Zhang, Si Li, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12834-12842
We present SOLID-Net, a neural network for spatially-varying outdoor lighting estimation from a single outdoor image for any 2D pixel location. Previous work has used a unified sky environment map to represent outdoor lighting. Instead, we generate spatially-varying local lighting environment maps by combining global sky environment map with warped image information according to geometric information estimated from intrinsics. As no outdoor dataset with image and local lighting ground truth is readily available, we introduce SOLID-Img dataset with physically-based rendered images and their corresponding intrinsic and lighting information. We train a deep neural network to regress intrinsic cues with physically-based constrains and use them to conduct global and local lightings estimation. Experiments on both synthetic and real datasets show that SOLID-Net significantly outperforms previous methods.
****
VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization
Seunghwan Choi, Sunghyun Park, Minsoo Lee, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14131-14140
The task of image-based virtual try-on aims to transfer a target clothing item o

nto the corresponding region of a person, which is commonly tackled by fitting t
he item to the desired body part and fusing the warped item with the person. Whi
le an increasing number of studies have been conducted, the resolution of synthe
sized images is still limited to low (e.g., 256x192), which acts as the critical
 limitation against satisfying online consumers. We argue that the limitation st
ems from several challenges: as the resolution increases, the artifacts in the m
isaligned areas between the warped clothes and the desired clothing regions beco
me noticeable in the final results; the architectures used in existing methods h
ave low performance in generating high-quality body parts and maintaining the te
xture sharpness of the clothes. To address the challenges, we propose a novel vi
rtual try-on method called VITON-HD that successfully synthesizes 1024x768 virtu
al try-on images. Specifically, we first prepare the segmentation map to guide o
ur virtual try-on synthesis, and then roughly fit the target clothing item to a
given person's body. Next, we propose ALIgnment-Aware Segment (ALIAS) normalizat
ion and ALIAS generator to handle the misaligned areas and preserve the details
of 1024x768 inputs. Through rigorous comparison with existing methods, we demons
trate that VITON-HD highly surpasses the baselines in terms of synthesized image
 quality both qualitatively and quantitatively.
****
Ultra-High-Definition Image Dehazing via Multi-Guided Bilateral Learning
Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Xiaobin Hu, Tao Wang, Fenglong Song, Xiu
yi Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2021, pp. 16185-16194
During the last couple of years, convolutional neural networks (CNNs) have achie
ved significant success in the single image dehazing task. Unfortunately, most e
xisting deep dehazing models have high computational complexity, which hinders t
heir application to high-resolution images, especially for UHD (ultra-high-defin
ition) or 4K resolution images. To address the problem, we propose a novel netwo
rk capable of real-time dehazing of 4K images on a single GPU, which consists of
 three deep CNNs. The first CNN extracts haze-relevant features at a reduced res
olution of the hazy input and then fits locally-affine models in the bilateral s
pace. Another CNN is used to learn multiple full-resolution guidance maps corres
ponding to the learned bilateral model. As a result, the feature maps with high-
frequency can be reconstructed by multi-guided bilateral upsampling. Finally, th
e third CNN fuses the high-quality feature maps into a dehazed image. In additio
n, we create a large-scale 4K image dehazing dataset to support the training and
 testing of compared models. Experimental results demonstrate that the proposed
algorithm performs favorably against the state-of-the-art dehazing approaches on
 various benchmarks.
****
RankDetNet: Delving Into Ranking Constraints for Object Detection
Ji Liu, Dong Li, Rongzhang Zheng, Lu Tian, Yi Shan; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 264-273
Modern object detection approaches cast detecting objects as optimizing two subt
asks of classification and localization simultaneously. Existing methods often l
earn the classification task by optimizing each proposal separately and neglect
the relationship among different proposals. Such detection paradigm also encount
ers the mismatch between classification and localization due to the inherent dis
crepancy of their optimization targets. In this work, we propose a ranking-based
 optimization algorithm for harmoniously learning to rank and localize proposals
 in lieu of the classification task. To this end, we comprehensively investigate
 three types of ranking constraints, i.e., global ranking, class-specific rankin
g and IoU-guided ranking losses. The global ranking loss encourages foreground s
amples to rank higher than background. The class-specific ranking loss ensures t
hat positive samples rank higher than negative ones for each specific class. The
 IoU-guided ranking loss aims to align each pair of confidence scores with the a
ssociated pair of IoU overlap between two positive samples of a specific class.
Our ranking constraints can sufficiently explore the relationships between sampl
es from three different perspectives. They are easy-to-implement, compatible wit
h mainstream detection frameworks and computation-free for inference. Experiment

s demonstrate that our RankDetNet consistently surpasses prior anchor-based and anchor-free baselines, e.g., improving RetinaNet baseline by 2.5% AP on the COCO test-dev set without bells and whistles. We also apply the proposed ranking constraints for 3D object detection and achieve improved performance, which further validates the superiority and generality of our method.
****

Back to the Feature: Learning Robust Camera Localization From Pixels To Pose
Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, Torsten Sattler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3247-3257
Camera pose estimation in known scenes is a 3D geometry task recently tackled by multiple learning algorithms. Many regress precise geometric quantities, like poses or 3D points, from an input image. This either fails to generalize to new viewpoints or ties the model parameters to a specific scene. In this paper, we go Back to the Feature: we argue that deep networks should focus on learning robust and invariant visual features, while the geometric estimation should be left to principled algorithms. We introduce PixLoc, a scene-agnostic neural network that estimates an accurate 6-DoF pose from an image and a 3D model. Our approach is based on the direct alignment of multiscale deep features, casting camera localization as metric learning. PixLoc learns strong data priors by end-to-end training from pixels to pose and exhibits exceptional generalization to new scenes by separating model parameters and scene geometry. The system can localize in large environments given coarse pose priors but also improve the accuracy of sparse feature matching by jointly refining keypoints and poses with little overhead. The code will be publicly available at github.com/cvg/pixloc.
****

Learning Parallel Dense Correspondence From Spatio-Temporal Descriptors for Efficient and Robust 4D Reconstruction
Jiapeng Tang, Dan Xu, Kui Jia, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6022-6031
This paper focuses on the task of 4D shape reconstruction from a sequence of point clouds. Despite the recent success achieved by extending deep implicit representations into 4D space, it is still a great challenge in two respects, i.e. how to design a flexible framework for learning robust spatio-temporal shape representations from 4D point clouds, and develop an efficient mechanism for capturing shape dynamics. In this work, we present a novel pipeline to learn a temporal evolution of the 3D human shape through spatially continuous transformation functions among cross-frame occupancy fields. The key idea is to parallelly establish the dense correspondence between predicted occupancy fields at different time steps via explicitly learning continuous displacement vector fields from robust spatio-temporal shape representations. Extensive comparisons against previous state-of-the-arts show the superior accuracy of our approach for 4D human reconstruction in the problems of 4D shape auto-encoding and completion, and a much faster network inference with about 8 times speedup demonstrates the significant efficiency of our approach.
****

Multi-Modal Fusion Transformer for End-to-End Autonomous Driving
Aditya Prakash, Kashyap Chitta, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7077-7087
How should representations from complementary sensors be integrated for autonomous driving? Geometry-based sensor fusion has shown great promise for perception tasks such as object detection and motion forecasting. However, for the actual driving task, the global context of the 3D scene is key, e.g. a change in traffic light state can affect the behavior of a vehicle geometrically distant from that traffic light. Geometry alone may therefore be insufficient for effectively fusing representations in end-to-end driving models. In this work, we demonstrate that imitation learning policies based on existing sensor fusion methods under-perform in the presence of a high density of dynamic agents and complex scenarios, which require global contextual reasoning, such as handling traffic oncoming f

rom multiple directions at uncontrolled intersections. Therefore, we propose TransFuser, a novel Multi-Modal Fusion Transformer, to integrate image and LiDAR representations using attention. We experimentally validate the efficacy of our approach in urban settings involving complex scenarios using the CARLA urban driving simulator. Our approach achieves state-of-the-art driving performance while reducing collisions by 76% compared to geometry-based fusion.
****
LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search
Bin Yan, Houwen Peng, Kan Wu, Dong Wang, Jianlong Fu, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15180-15189
Object tracking has achieved significant progress over the past few years. However, state-of-the-art trackers become increasingly heavy and expensive, which limits their deployments in resource-constrained applications. In this work, we present LightTrack, which uses neural architecture search (NAS) to design more lightweight and efficient object trackers. Comprehensive experiments show that our LightTrack is effective. It can find trackers that achieve superior performance compared to handcrafted SOTA trackers, such as SiamRPN++ and Ocean, while using much fewer model Flops and parameters. Moreover, when deployed on resource-constrained mobile chipsets, the discovered trackers run much faster. For example, on Snapdragon 845 Adreno GPU, LightTrack runs 12x faster than Ocean, while using 13x fewer parameters and 38x fewer Flops. Such improvements might narrow the gap between academic models and industrial deployments in object tracking task. LightTrack is released at here.
****
Unsupervised Disentanglement of Linear-Encoded Facial Semantics
Yutong Zheng, Yu-Kai Huang, Ran Tao, Zhiqiang Shen, Marios Savvides; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3917-3926
We propose a method to disentangle linear-encoded facial semantics from StyleGAN without external supervision. The method derives from linear regression and sparse representation learning concepts to make the disentangled latent representations easily interpreted as well. We start by coupling StyleGAN with a stabilized 3D deformable facial reconstruction method to decompose single-view GAN generations into multiple semantics. Latent representations are then extracted to capture interpretable facial semantics. In this work, we make it possible to get rid of labels for disentangling meaningful facial semantics. Also, we demonstrate that the guided extrapolation along the disentangled representations can help with data augmentation, which sheds light on handling unbalanced data. Finally, we provide an analysis of our learned localized facial representations and illustrate that the semantic information is encoded, which surprisingly complies with human intuition. The overall unsupervised design brings more flexibility to representation learning in the wild.
****
Learning Position and Target Consistency for Memory-Based Video Object Segmentation
Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, Rong Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4144-4154
This paper studies the problem of semi-supervised video object segmentation(VOS). Multiple works have shown that memory-based approaches can be effective for video object segmentation. They are mostly based on pixel-level matching, both spatially and temporally. The main shortcoming of memory-based approaches is that they do not take into account the sequential order among frames and do not exploit object-level knowledge from the target. To address this limitation, we propose to learn position and target consistency framework for memory-based video object segmentation, termed as LCM. It applies the memory mechanism to retrieve pixels globally, and meanwhile learns position consistency for more reliable segmentation. The learned location response promotes a better discrimination between tar

get and distractors. Besides, LCM introduces an object-level relationship from the target to maintain target consistency, making LCM more robust to error drifting. Experiments show that our LCM achieves state-of-the-art performance on both DAVIS and Youtube-VOS benchmark. And we rank the 1st in the DAVIS 2020 challenge semi-supervised VOS task.
****
Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation
Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12414-12424
Self-training is a competitive approach in domain adaptive segmentation, which trains the network with the pseudo labels on the target domain. However inevitably, the pseudo labels are noisy and the target features are dispersed due to the discrepancy between source and target domains. In this paper, we rely on representative prototypes, the feature centroids of classes, to address the two issues for unsupervised domain adaptation. In particular, we take one step further and exploit the feature distances from prototypes that provide richer information than mere prototypes. Specifically, we use it to estimate the likelihood of pseudo labels to facilitate online correction in the course of training. Meanwhile, we align the prototypical assignments based on relative feature distances for two different views of the same target, producing a more compact target feature space. Moreover, we find that distilling the already learned knowledge to a self-supervised pretrained model further boosts the performance. Our method shows tremendous performance advantage over state-of-the-art methods.
****
Deep Denoising of Flash and No-Flash Pairs for Photography in Low-Light Environments
Zhihao Xia, Michael Gharbi, Federico Perazzi, Kalyan Sunkavalli, Ayan Chakrabarti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2063-2072
We introduce a neural network-based method to denoise pairs of images taken in quick succession in low-light environments, with and without a flash. Our goal is to produce a high-quality rendering of the scene that preserves the color and mood from the ambient illumination of the noisy no-flash image, while recovering surface texture and detail revealed by the flash. Our network outputs a gain map and a field of kernels, the latter obtained by linearly mixing elements of a per-image low-rank kernel basis. We first apply the kernel field to the no-flash image, and then multiply the result with the gain map to create the final output. We show our network effectively learns to produce high-quality images by combining a smoothed out estimate of the scene's ambient appearance from the no-flash image, with high-frequency albedo details extracted from the flash input. Our experiments show significant improvements over alternative captures without a flash, and baseline denoisers that use flash no-flash pairs. In particular, our method produces images that are both noise-free and contain accurate ambient colors without the sharp shadows or strong specular highlights visible in the flash image.
****
Transformer Interpretability Beyond Attention Visualization
Hila Chefer, Shir Gur, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782-791
Self-attention techniques, and specifically Transformers, are dominating the field of text processing and are becoming increasingly popular in computer vision classification tasks. In order to visualize the parts of the image that led to a certain classification, existing methods either rely on the obtained attention maps or employ heuristic propagation along the attention graph. In this work, we propose a novel way to compute relevancy for Transformer networks. The method assigns local relevance based on the Deep Taylor Decomposition principle and then propagates these relevancy scores through the layers. This propagation involves attention layers and skip connections, which challenge existing methods. Our sol

ution is based on a specific formulation that is shown to maintain the total relevancy across layers. We benchmark our method on very recent visual Transformer networks, as well as on a text classification problem, and demonstrate a clear advantage over the existing explainability methods.
****
Unsupervised Learning for Robust Fitting: A Reinforcement Learning Approach
Giang Truong, Huu Le, David Suter, Erchuan Zhang, Syed Zulqarnain Gilani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10348-10357
Robust model fitting is a core algorithm in a large number of computer vision applications. Solving this problem efficiently for highly contaminated datasets is, however, still challenging due to its underlying computational complexity. Recent attention has been focused on learning-based algorithms. However, most approaches are supervised (which require a large amount of labelled training data). In this paper, we introduce a novel unsupervised learning framework that learns to directly solve robust model fitting. Unlike other methods, our work is agnostic to the underlying input features, and can be easily generalized to a wide variety of LP-type problems with quasi-convex residuals. We empirically show that our method outperforms existing unsupervised learning approaches, and achieves competitive results compared to traditional methods on several important computer vision problems.
****
Unsupervised Real-World Image Super Resolution via Domain-Distance Aware Training
Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, Hengjie Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13385-13394
These days, unsupervised super-resolution (SR) is soaring due to its practical and promising potential in real scenarios. The philosophy of off-the-shelf approaches lies in the augmentation of unpaired data, i.e. first generating synthetic low-resolution (LR) images Y^g corresponding to real-world high-resolution (HR) images X^r in the real-world LR domain Y^r, and then utilizing the pseudo pairs Y^g, X^r for training in a supervised manner. Unfortunately, since image translation itself is an extremely challenging task, the SR performance of these approaches is severely limited by the domain gap between generated synthetic LR images and real LR images. In this paper, we propose a novel domain-distance aware super-resolution (DASR) approach for unsupervised real-world image SR. The domain gap between training data (e.g. Y^g) and testing data (e.g. Y^r) is addressed with our domain-gap aware training and domain-distance weighted supervision strategies. Domain-gap aware training takes additional benefit from real data in the target domain while domain-distance weighted supervision brings forward the more rational use of labeled source domain data. The proposed method is validated on synthetic and real datasets and the experimental results show that DASR consistently outperforms state-of-the-art unsupervised SR approaches in generating SR outputs with more realistic and natural textures. Codes are available at https://github.com/ShuhangGu/DASR.
****
Learning to Track Instances without Video Annotations
Yang Fu, Sifei Liu, Umar Iqbal, Shalini De Mello, Humphrey Shi, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8680-8689
Tracking segmentation masks of multiple instances has been intensively studied, but still faces two fundamental challenges: 1) the requirement of large-scale, frame-wise annotation, and 2) the complexity of two-stage approaches. To resolve these challenges, we introduce a novel semi-supervised framework by learning instance tracking networks with only a labeled image dataset and unlabeled video sequences. With an instance contrastive objective, we learn an embedding to discriminate each instance from the others. We show that even when only trained with images, the learned feature representation is robust to instance appearance variations, and is thus able to track objects steadily across frames. We further enha

nce the tracking capability of the embedding by learning correspondence from unl
abeled videos in a self-supervised manner. In addition, we integrate this module
 into single-stage instance segmentation and pose estimation frameworks, which s
ignificantly reduce the computational complexity of tracking compared to two-sta
ge networks. We conduct experiments on the YouTube-VIS and PoseTrack datasets. W
ithout any video annotation efforts, our proposed method can achieve comparable
or even better performance than most fully-supervised methods.
****
Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination
Xudong Wang, Ziwei Liu, Stella X. Yu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12586-12595
Unsupervised feature learning has made great strides with contrastive learning b
ased on instance discrimination and invariant mapping, as benchmarked on curated
 class-balanced datasets. However, natural data could be highly correlated and l
ong-tail distributed. Natural between-instance similarity conflicts with the pre
sumed instance distinction, causing unstable training and poor performance. Our
idea is to discover and integrate between-instance similarity into contrastive l
earning, not directly by instance grouping, but by cross-level discrimination (C
LD) between instances and local instance groups. While invariant mapping of each
 instance is imposed by attraction within its augmented views, between-instance
similarity emerges from common repulsion against instance groups. Our batch-wise
 and cross-view comparisons also greatly improve the positive/negative sample ra
tio of contrastive learning and achieve better invariant mapping. To effect both
 grouping and discrimination objectives, we impose them on features separately d
erived from a shared representation. In addition, we propose normalized projecti
on heads and unsupervised hyper-parameter tuning for the first time. Our extensi
ve experimentation demonstrates that CLD is a lean and powerful add-on to existi
ng methods (e.g., NPID, MoCo, InfoMin, BYOL) on highly correlated, long-tail, or
 balanced datasets. It not only achieves new state-of-the-art on self-supervisio
n, semi-supervision, and transfer learning benchmarks, but also beats MoCo v2 an
d SimCLR on every reported performance attained with a much larger compute. CLD
effectively extends unsupervised learning to natural data and brings it closer t
o real-world applications.
****
Representation Learning via Global Temporal Alignment and Cycle-Consistency
Isma Hadji, Konstantinos G. Derpanis, Allan D. Jepson; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11068
-11077
We introduce a weakly supervised method for representation learning based on ali
gning temporal sequences (e.g., videos) of the same process (e.g., human action)
. The main idea is to use the global temporal ordering of latent correspondences
 across sequence pairs as a supervisory signal. In particular, we propose a loss
 based on scoring the optimal sequence alignment to train an embedding network.
Our loss is based on a novel probabilistic path finding view of dynamic time war
ping (DTW) that contains the following three key features: (i) the local path ro
uting decisions are contrastive and differentiable, (ii) pairwise distances are
cast as probabilities that are contrastive as well, and (iii) our formulation na
turally admits a global cycle consistency loss that verifies correspondences. Fo
r evaluation, we consider the tasks of fine-grained action classification, few s
hot learning, and video synchronization. We report significant performance incre
ases over previous methods. In addition, we report two applications of our tempo
ral alignment framework, namely 3D pose reconstruction and fine-grained audio/vi
sual retrieval.
****
Personalized Outfit Recommendation With Learnable Anchors
Zhi Lu, Yang Hu, Yan Chen, Bing Zeng; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12722-12731
The multimedia community has recently seen a tremendous surge of interest in the
 fashion recommendation problem. A lot of efforts have been made to model the co
mpatibility between fashion items. Some have also studied users' personal prefer

ences for the outfits. There is, however, another difficulty in the task that hasn't been dealt with carefully by previous work. Users that are new to the system usually only have several (less than 5) outfits available for learning. With such a limited number of training examples, it is challenging to model the user's preferences reliably. In this work, we propose a new solution for personalized outfit recommendation that is capable of handling this case. We use a stacked self-attention mechanism to model the high-order interactions among the items. We then embed the items in an outfit into a single compact representation within the outfit space. To accommodate the variety of users' preferences, we characterize each user with a set of anchors, i.e. a group of learnable latent vectors in the outfit space that are the representatives of the outfits the user likes. We also learn a set of general anchors to model the general preference shared by all users. Based on this representation of the outfits and the users, we propose a simple but effective strategy for the new user profiling tasks. Extensive experiments on large scale real-world datasets demonstrate the performance of our proposed method.
****
When Age-Invariant Face Recognition Meets Face Age Synthesis: A Multi-Task Learning Framework
Zhizhong Huang, Junping Zhang, Hongming Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7282-7291
To minimize the effects of age variation in face recognition, previous work either extracts identity-related discriminative features by minimizing the correlation between identity- and age-related features, called age-invariant face recognition (AIFR), or removes age variation by transforming the faces of different age groups into the same age group, called face age synthesis (FAS); however, the former lacks visual results for model interpretation while the latter suffers from artifacts compromising downstream recognition. Therefore, this paper proposes a unified, multi-task framework to jointly handle these two tasks, termed MTLFace, which can learn age-invariant identity-related representation while achieving pleasing face synthesis. Specifically, we first decompose the mixed face features into two uncorrelated components---identity- and age-related features---through an attention mechanism, and then decorrelate these two components using multi-task training and continuous domain adaption. In contrast to the conventional one-hot encoding that achieves group-level FAS, we propose a novel identity conditional module to achieve identity-level FAS, with a weight-sharing strategy to improve the age smoothness of synthesized faces. In addition, we collect and release a large cross-age face dataset with age and gender annotations to advance AIFR and FAS. Extensive experiments on five benchmark cross-age datasets demonstrate the superior performance of our proposed MTLFace over state-of-the-art methods for AIFR and FAS. We further validate MTLFace on two popular general face recognition datasets, showing competitive performance for face recognition in the wild. The source code and dataset are available at https://github.com/Hzzone/MTLFace.
****
Learning Dynamics via Graph Neural Networks for Human Pose Estimation and Tracking
Yiding Yang, Zhou Ren, Haoxiang Li, Chunluan Zhou, Xinchao Wang, Gang Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8074-8084
Multi-person pose estimation and tracking serve as crucial steps for video understanding. Most state-of-the-art approaches rely on first estimating poses in each frame and only then implementing data association and refinement. Despite the promising results achieved, such a strategy is inevitably prone to missed detections especially in heavily-cluttered scenes, since this tracking-by-detection paradigm is, by nature, largely dependent on visual evidences that are absent in the case of occlusion. In this paper, we propose a novel online approach to learning the pose dynamics, which are independent of pose detections in current fame, and hence may serve as a robust estimation even in challenging scenarios including occlusion. Specifically, we derive this prediction of dynamics through a gra

ph neural network (GNN) that explicitly accounts for both spatial-temporal and v isual information. It takes as input the historical pose tracklets and directly predicts the corresponding poses in the following frame for each tracklet. The p redicted poses will then be aggregated with the detected poses, if any, at the s ame frame so as to produce the final pose, potentially recovering the occluded j oints missed by the estimator. Experiments on PoseTrack 2017 and PoseTrack 2018 datasets demonstrate that the proposed method achieves results superior to the s tate of the art on both human pose estimation and tracking tasks.
****
Smoothing the Disentangled Latent Style Space for Unsupervised Image-to-Image Tr anslation
Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, Marco De Nadai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10785-10794
Image-to-Image (I2I) multi-domain translation models are usually evaluated also using the quality of their semantic interpolation results. However, state-of-the -art models frequently show abrupt changes in the image appearance during interp olation, and usually perform poorly in interpolations across domains. In this pa per, we propose a new training protocol based on three specific losses which hel p a translation network to learn a smooth and disentangled latent style space in which: 1) Both intra- and inter-domain interpolations correspond to gradual cha nges in the generated images and 2) The content of the source image is better pr eserved during the translation. Moreover, we propose a novel evaluation metric t o properly measure the smoothness of latent style space of I2I translation model s. The proposed method can be plugged in existing translation approaches, and ou r extensive experiments on different datasets show that it can significantly boo st the quality of the generated images and the graduality of the interpolations.
****
Robust Instance Segmentation Through Reasoning About Multi-Object Occlusion
Xiaoding Yuan, Adam Kortylewski, Yihong Sun, Alan Yuille; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11 141-11150
Analyzing complex scenes with Deep Neural Networks is a challenging task, partic ularly when images contain multiple objects that partially occlude each other. E xisting approaches to image analysis mostly process objects independently and do not take into account the relative occlusion of nearby objects. In this paper, we propose a deep network for multi-object instance segmentation that is robust to occlusion and can be trained from bounding box supervision only. Our work bui lds on Compositional Networks, which learn a generative model of neural feature activations to locate occluders and to classify objects based on their non-occlu ded parts. We extend their generative model to include multiple objects and intr oduce a framework for efficient inference in challenging occlusion scenarios. In particular, we obtain feed-forward predictions of the object classes and their instance and occluder segmentations. We introduce an Occlusion Reasoning Module (ORM) that locates erroneous segmentations and estimates the occlusion order to correct them. The improved segmentation masks are, in turn, integrated into the network in a top-down manner to improve the image classification. Our experiment s on the KITTI INStance dataset (KINS) and a synthetic occlusion dataset demonst rate the effectiveness and robustness of our model at multi-object instance segm entation under occlusion. Code is publically available at https://github.com/XD7 479/Multi-Object-Occlusion.
****
Architectural Adversarial Robustness: The Case for Deep Pursuit
George Cazenavette, Calvin Murdock, Simon Lucey; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7150-7158
Despite their unmatched performance, deep neural networks remain susceptible to targeted attacks by nearly imperceptible levels of adversarial noise. While the underlying cause of this sensitivity is not well understood, theoretical analyse s can be simplified by reframing each layer of a feed-forward network as an appr oximate solution to a sparse coding problem. Iterative solutions using basis pur

suit are theoretically more stable and have improved adversarial robustness. However, cascading layer-wise pursuit implementations suffer from error accumulation in deeper networks. In contrast, our new method of deep pursuit approximates the activations of all layers as a single global optimization problem, allowing us to consider deeper, real-world architectures with skip connections such as residual networks. Experimentally, our approach demonstrates improved robustness to adversarial noise.
****
Multi-Scale Aligned Distillation for Low-Resolution Detection
Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14443-14453
In instance-level detection tasks (e.g., object detection), reducing input resolution is an easy option to improve runtime efficiency. However, this option severely hurts the detection performance. This paper focuses on boosting the performance of a low-resolution model, by distilling knowledge from a high/multi-resolution model. We first identify the challenge of applying knowledge distillation to teacher and student networks that act on different input resolutions. To tackle the challenge, we explore the idea of spatially aligning feature maps between models of different input resolutions, by shifting the position of the feature pyramid structure. With the alignment idea, we introduce aligned multi-scale training to train a multi-scale teacher that can distill its knowledge seamlessly to a low-resolution student. Furthermore, we propose cross feature-level fusion to dynamically fuse the multi-resolution features of the same teacher, to better guide the student. On several instance-level detection tasks and datasets, the low-resolution models trained via our approach perform competitively with high-resolution models trained via conventional multi-scale training, while outperforming the latter's low-resolution models by 2.1% to 3.6% in mAP.
****
Deep Active Surface Models
Udaranga Wickramasinghe, Pascal Fua, Graham Knott; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11652-11661
Active Surface Models have a long history of being useful to model complex 3D surfaces. But only Active Contours have been used in conjunction with deep networks, and then only to produce the data term as well as meta-parameter maps controlling them. In this paper, we advocate a much tighter integration. We introduce layers that implement them that can be integrated seamlessly into Graph Convolutional Networks to enforce sophisticated smoothness priors at an acceptable computational cost. We will show that the resulting Deep Active Surface Models outperform equivalent architectures that use traditional regularization loss terms to impose smoothness priors for 3D surface reconstruction from 2D images and for 3D volume segmentation.
****
Can We Characterize Tasks Without Labels or Features?
Bram Wallace, Ziyang Wu, Bharath Hariharan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1245-1254
The problem of expert model selection deals with choosing the appropriate pretrained network ("expert") to transfer to a target task. Methods, however, generally depend on two separate assumptions: the presence of labeled images and access to powerful "probe" networks that yield useful features. In this work, we demonstrate the current reliance on both of these aspects and develop algorithms to operate when either of these assumptions fail. In the unlabeled case, we show that pseudolabels from the probe network provide discriminative enough gradients to perform nearly-equal task selection even when the probe network is trained on imagery unrelated to the tasks. To compute the embedding with no probe network at all, we introduce the Task Tangent Kernel (TTK) which uses a kernelized distance across multiple random networks to achieve performance over double that of other methods with randomly initialized models. Code is available at https://github.com/BramSW/task_characterization_cvpr_2021/.

****
Scene Essence
Jiayan Qiu, Yiding Yang, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8322-8333
What scene elements, if any, are indispensable for recognizing a scene? We strive to answer this question through the lens of an end-to-end learning scheme. Our goal is to identify a collection of such pivotal elements, which we term as Scene Essence, to be those that would alter scene recognition if taken out from the scene. To this end, we devise a novel approach that learns to partition the scene objects into two groups, essential ones and minor ones, under the supervision that if only the essential ones are kept while the minor ones are erased in the input image, a scene recognizer would preserve its original prediction. Specifically, we introduce a learnable graph neural network (GNN) for labelling scene objects, based on which the minor ones are wiped off by an off-the-shelf image inpainter. The features of the inpainted image derived in this way, together with those learned from the GNN with the minor-object nodes pruned, are expected to fool the scene discriminator. Both subjective and objective evaluations on Places365, SUN397, and MIT67 datasets demonstrate that, the learned Scene Essence yields a visually plausible image that convincingly retains the original scene category.
****
Visual Room Rearrangement
Luca Weihs, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5922-5931
There has been a significant recent progress in the field of Embodied AI with researchers developing models and algorithms enabling embodied agents to navigate and interact within completely unseen environments. In this paper, we propose a new dataset and baseline models for the task of Rearrangement. We particularly focus on the task of Room Rearrangement: an agent begins by exploring a room and recording objects' initial configurations. We then remove the agent and change the poses and states (e.g., open/closed) of some objects in the room. The agent must restore the initial configurations of all objects in the room. Our dataset, named RoomR, includes 6,000 distinct rearrangement settings involving 72 different object types in 120 scenes. Our experiments show that solving this challenging interactive task that involves navigation and object interaction is beyond the capabilities of the current state-of-the-art techniques for embodied tasks and we are still very far from achieving perfect performance on these types of tasks.
****
VDSM: Unsupervised Video Disentanglement With State-Space Modeling and Deep Mixtures of Experts
Matthew J. Vowels, Necati Cihan Camgoz, Richard Bowden; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8176-8186
Disentangled representations support a range of downstream tasks including causal reasoning, generative modeling, and fair machine learning. Unfortunately, disentanglement has been shown to be impossible without the incorporation of supervision or inductive bias. Given that supervision is often expensive or infeasible to acquire, we choose to incorporate structural inductive bias and present an unsupervised, deep State-Space-Model for Video Disentanglement (VDSM). The model disentangles latent time-varying and dynamic factors via the incorporation of hierarchical structure with a dynamic prior and a Mixture of Experts decoder. VDSM learns separate disentangled representations for the identity of the object or person in the video, and for the action being performed. We evaluate VDSM across a range of qualitative and quantitative tasks including identity and dynamics transfer, sequence generation, Frechet Inception Distance, and factor classification. VDSM achieves state-of-the-art performance and exceeds adversarial methods, even when the methods use additional supervision.

****
Rotation-Only Bundle Adjustment
Seong Hun Lee, Javier Civera; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 424-433
We propose a novel method for estimating the global rotations of the cameras independently of their positions and the scene structure. When two calibrated cameras observe five or more of the same points, their relative rotation can be recovered independently of the translation. We extend this idea to multiple views, thereby decoupling the rotation estimation from the translation and structure estimation. Our approach provides several benefits such as complete immunity to inaccurate translations and structure, and the accuracy improvement when used with rotation averaging. We perform extensive evaluations on both synthetic and real datasets, demonstrating consistent and significant gains in accuracy when used with the state-of-the-art rotation averaging method.
****
Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting With Their Explanations
Wolfgang Stammer, Patrick Schramowski, Kristian Kersting; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3619-3629
Most explanation methods in deep learning map importance estimates for a model's prediction back to the original input space. These "visual" explanations are often insufficient, as the model's actual concept remains elusive. Moreover, without insights into the model's semantic concept, it is difficult --if not impossible-- to intervene on the model's behavior via its explanations, called Explanatory Interactive Learning. Consequently, we propose to intervene on a Neuro-Symbolic scene representation, which allows one to revise the model on the semantic level, e.g. "never focus on the color to make your decision". We compiled a novel con-founded visual scene data set, the CLEVR-Hans data set,capturing complex compositions of different objects. The results of our experiments on CLEVR-Hans demonstrate that our semantic explanations, i.e. compositional explanations at a per-object level, can identify confounders that are not identifiable using "visual" explanations only. More importantly, feedback on this semantic level makes it possible to revise the model from focusing on these factors.
****
Polygonal Point Set Tracking
Gunhee Nam, Miran Heo, Seoung Wug Oh, Joon-Young Lee, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5569-5578
In this paper, we propose a novel learning-based polygonal point set tracking method. Compared to existing video object segmentation (VOS) methods that propagate pixel-wise object mask information, we propagate a polygonal point set over frames. Specifically, the set is defined as a subset of points in the target contour, and our goal is to track corresponding points on the target contour. Those outputs enable us to apply various visual effects such as motion tracking, part deformation, and texture mapping. To this end, we propose a new method to track the corresponding points between frames by the global-local alignment with delicately designed losses and regularization terms. We also introduce a novel learning strategy using synthetic and VOS datasets that makes it possible to tackle the problem without developing the point correspondence dataset. Since the existing datasets are not suitable to validate our method, we build a new polygonal point set tracking dataset and demonstrate the superior performance of our method over the baselines and existing contour-based VOS methods. In addition, we present visual-effects applications of our method on part distortion and text mapping.
****
Deformed Implicit Field: Modeling 3D Shapes With Learned Dense Correspondence
Yu Deng, Jiaolong Yang, Xin Tong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10286-10296
We propose a novel Deformed Implicit Field (DIF) representation for modeling 3D shapes of a category and generating dense correspondences among shapes. With DIF

, a 3D shape is represented by a template implicit field shared across the categ
ory, together with a 3D deformation field and a correction field dedicated for e
ach shape instance. Shape correspondences can be easily established using their
deformation fields. Our neural network, dubbed DIF-Net, jointly learns a shape l
atent space and these fields for 3D objects belonging to a category without usin
g any correspondence or part label. The learned DIF-Net can also provides reliab
le correspondence uncertainty measurement reflecting shape structure discrepancy
. Experiments show that DIF-Net not only produces high-fidelity 3D shapes but al
so builds high-quality dense correspondences across different shapes. We also de
monstrate several applications such as texture transfer and shape editing, where
 our method achieves compelling results that cannot be achieved by previous meth
ods.
****
Verifiability and Predictability: Interpreting Utilities of Network Architecture
s for Point Cloud Processing
Wen Shen, Zhihua Wei, Shikun Huang, Binbin Zhang, Panyue Chen, Ping Zhao, Quansh
i Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2021, pp. 10703-10712
In this paper, we diagnose deep neural networks for 3D point cloud processing to
 explore utilities of different network architectures. We propose a number of hy
potheses on the effects of specific network architectures on the representation
capacity of DNNs. In order to prove the hypotheses, we design five metrics to di
agnose various types of DNNs from the following perspectives, information discar
ding, information concentration, rotation robustness, adversarial robustness, an
d neighborhood inconsistency. We conduct comparative studies based on such metri
cs to verify the hypotheses. We further use the verified hypotheses to revise ar
chitectures of existing DNNs and improve their utilities. Experiments demonstrat
e the effectiveness of our method. The code will be released when this paper is
accepted.
****
Tracking Pedestrian Heads in Dense Crowd
Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, Julien Pettre; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2021, pp. 3865-3875
Tracking humans in crowded video sequences is an important constituent of visual
 scene understanding. Increasing crowd density challenges visibility of humans,
limiting the scalability of existing pedestrian trackers to higher crowd densiti
es. For that reason, we propose to revitalize head tracking with Crowd of Heads
Dataset (CroHD), consisting of 9 sequences of 11,463 frames with over 2,276,838
heads and 5,230 tracks annotated in diverse scenes. For evaluation, we proposed
a new metric, IDEucl, to measure an algorithm's efficacy in preserving a unique
identity for the longest stretch in image coordinate space, thus building a corr
espondence between pedestrian crowd motion and the performance of a tracking alg
orithm. Moreover, we also propose a new head detector, HeadHunter, which is desi
gned for small head detection in crowded scenes. We extend HeadHunter with a Par
ticle Filter and a color histogram based re-identification module for head track
ing. To establish this as a strong baseline, we compare our tracker with existin
g state-of-the-art pedestrian trackers on CroHD and demonstrate superiority, esp
ecially in identity preserving tracking metrics. With a light-weight head detect
or and a tracker which is efficient at identity preservation, we believe our con
tributions will serve useful in advancement of pedestrian tracking in dense crow
ds. We make our dataset, code and models publicly available at https://project.i
nria.fr/crowdscience/project/dense-crowd-head-tracking/.
****
Neural Splines: Fitting 3D Surfaces With Infinitely-Wide Neural Networks
Francis Williams, Matthew Trager, Joan Bruna, Denis Zorin; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9
949-9958
We present Neural Splines, a technique for 3D surface reconstruction that is bas
ed on random feature kernels arising from infinitely-wide shallow ReLU networks.

Our method achieves state-of-the-art results, outperforming recent neural network-based techniques and widely used Poisson Surface Reconstruction (which, as we demonstrate, can also be viewed as a type of kernel method). Because our approach is based on a simple kernel formulation, it is easy to analyze and can be accelerated by general techniques designed for kernel-based learning. We provide explicit analytical expressions for our kernel and argue that our formulation can be seen as a generalization of cubic spline interpolation to higher dimensions. In particular, the RKHS norm associated with Neural Splines biases toward smooth interpolants.
****
Alpha-Refine: Boosting Tracking Performance by Precise Bounding Box Estimation
Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, Xiaoyun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5289-5298
Visual object tracking aims to precisely estimate the bounding box for the given target, which is a challenging problem due to factors such as deformation and occlusion. Many recent trackers adopt the multiple-stage tracking strategy to improve the quality of bounding box estimation. These methods first coarsely locate the target and then refine the initial prediction in the following stages. However, existing approaches still suffer from limited precision, and the coupling of different stages severely restricts the method's transferability. This work proposes a novel, flexible, and accurate refinement module called Alpha-Refine (AR), which can significantly improve the base trackers' box estimation quality. By exploring a series of design options, we conclude that the key to successful refinement is extracting and maintaining detailed spatial information as much as possible. Following this principle, Alpha-Refine adopts a pixel-wise correlation, a corner prediction head, and an auxiliary mask head as the core components. Comprehensive experiments on TrackingNet, LaSOT, GOT-10K, and VOT2020 benchmarks with multiple base trackers show that our approach significantly improves the base trackers' performance with little extra latency. The proposed Alpha-Refine method leads to a series of strengthened trackers, among which the ARSiamRPN (AR strengthened SiamRPNpp) and the ARDiMP50 (ARstrengthened DiMP50) achieve good efficiency-precision trade-off, while the ARDiMPsuper (AR strengthened DiMP-super) achieves very competitive performance at a real-time speed. Code and pretrained models are available at https://github.com/MasterBin-IIAU/AlphaRefine.
****
Adaptive Cross-Modal Prototypes for Cross-Domain Visual-Language Retrieval
Yang Liu, Qingchao Chen, Samuel Albanie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14954-14964
In this paper, we study the task of visual-text retrieval in the highly practical setting in which labelled visual data with paired text descriptions are available in one domain (the "source"), but only unlabelled visual data (without text descriptions) are available in the domain of interest (the "target"). We propose the ADAPTIVE CROSS-MODAL PROTOTYPES framework which seeks to enable target domain retrieval by learning cross-modal visual-text representations while minimising both uni-modal and cross-modal distribution shift across the source and target domains. Our approach is built upon two key ideas: first, we encode the inductive bias that the learned cross-modal representations should be compositional with respect to concepts in each modality--this is achieved through clustering pretrained uni-modal features across each domain and designing a careful regularisation scheme to preserve the resulting structure. Second, we employ mutual information maximisation between cross-modal representations in the source and target domains during learning--this provides a mechanism that preserves commonalities between the domains while discarding signal in each that cannot be inferred from the other. We showcase our approach for the task of cross-domain visual-text retrieval, outperforming existing approaches for both images and videos.
****
Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts
Soravit Changpinyo, Piyush Sharma, Nan Ding, Radu Soricut; Proceedings of the IE

EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3558-3568
The availability of large-scale image captioning and visual question answering datasets has contributed significantly to recent successes in vision-and-language pre-training. However, these datasets are often collected with overrestrictive requirements inherited from their original target tasks (e.g., image caption generation), which limit the resulting dataset scale and diversity. We take a step further in pushing the limits of vision-and-language pre-training data by relaxing the data collection pipeline used in Conceptual Captions 3M (CC3M) [Sharma et al. 2018] and introduce the Conceptual 12M (CC12M), a dataset with 12 million image-text pairs specifically meant to be used for vision-and-language pre-training. We perform an analysis of this dataset and benchmark its effectiveness against CC3M on multiple downstream tasks with an emphasis on long-tail visual recognition. Our results clearly illustrate the benefit of scaling up pre-training data for vision-and-language tasks, as indicated by the new state-of-the-art results on both the nocaps and Conceptual Captions benchmarks.
****
SetVAE: Learning Hierarchical Composition for Generative Modeling of Set-Structured Data
Jinwoo Kim, Jaehoon Yoo, Juho Lee, Seunghoon Hong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15059-15068
Generative modeling of set-structured data, such as point clouds, requires reasoning over local and global structures at various scales. However, adopting multi-scale frameworks for ordinary sequential data to a set-structured data is nontrivial as it should be invariant to the permutation of its elements. In this paper, we propose SetVAE, a hierarchical variational autoencoder for sets. Motivated by recent progress in set encoding, we build SetVAE upon attentive modules that first partition the set and project the partition back to the original cardinality. Exploiting this module, our hierarchical VAE learns latent variables at multiple scales, capturing coarse-to-fine dependency of the set elements while achieving permutation invariance. We evaluate our model on point cloud generation task and achieve competitive performance to the prior arts with substantially smaller model capacity. We qualitatively demonstrate that our model generalizes to unseen set sizes and learns interesting subset relations without supervision. Our implementation is available at https://github.com/jw9730/setvae.
****
Few-Shot 3D Point Cloud Semantic Segmentation
Na Zhao, Tat-Seng Chua, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8873-8882
Many existing approaches for 3D point cloud semantic segmentation are fully supervised. These fully supervised approaches heavily rely on large amounts of labeled training data that are difficult to obtain and cannot segment new classes after training. To mitigate these limitations, we propose a novel attention-aware multi-prototype transductive few-shot point cloud semantic segmentation method to segment new classes given a few labeled examples. Specifically, each class is represented by multiple prototypes to model the complex data distribution of labeled points. Subsequently, we employ a transductive label propagation method to exploit the affinities between labeled multi-prototypes and unlabeled points, and among the unlabeled points. Furthermore, we design an attention-aware multi-level feature learning network to learn the discriminative features that capture the geometric dependencies and semantic correlations between points. Our proposed method shows significant and consistent improvements compared to baselines in different few-shot point cloud semantic segmentation settings (i.e., 2/3-way 1/5-shot) on two benchmark datasets. Our code is available at https://github.com/Na-Z/attMPTI.
****
CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching
Zhelun Shen, Yuchao Dai, Zhibo Rao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13906-13915

Recently, the ever-increasing capacity of large-scale annotated datasets has led to profound progress in stereo matching. However, most of these successes are limited to a specific dataset and cannot generalize well to other datasets. The main difficulties lie in the large domain differences and unbalanced disparity distribution across a variety of datasets, which greatly limit the real-world applicability of current deep stereo matching models. In this paper, we propose CFNet, a Cascade and Fused cost volume based network to improve the robustness of the stereo matching network. First, we propose a fused cost volume representation to deal with the large domain difference. By fusing multiple low-resolution dense cost volumes to enlarge the receptive field, we can extract robust structural representations for initial disparity estimation. Second, we propose a cascade cost volume representation to alleviate the unbalanced disparity distribution. Specifically, we employ a variance-based uncertainty estimation to adaptively adjust the next stage disparity search space, in this way driving the network progressively prune out the space of unlikely correspondences. By iteratively narrowing down the disparity search space and improving the cost volume resolution, the disparity estimation is gradually refined in a coarse-to-fine manner. When trained on the same training images and evaluated on KITTI, ETH3D, and Middlebury datasets with the fixed model parameters and hyperparameters, our proposed method achieves the state-of-the-art overall performance and obtains the 1st place on the stereo task of Robust Vision Challenge 2020. The code will be available at https://github.com/gallenszl/CFNet.
****
Adaptive Consistency Prior Based Deep Network for Image Denoising
Chao Ren, Xiaohai He, Chuncheng Wang, Zhibo Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8596-8606
Recent studies have shown that deep networks can achieve promising results for image denoising. However, how to simultaneously incorporate the valuable achievements of traditional methods into the network design and improve network interpretability is still an open problem. To solve this problem, we propose a novel model-based denoising method to inform the design of our denoising network. First, by introducing a non-linear filtering operator, a reliability matrix, and a high-dimensional feature transformation function into the traditional consistency prior, we propose a novel adaptive consistency prior (ACP). Second, by incorporating the ACP term into the maximum a posteriori framework, a model-based denoising method is proposed. This method is further used to inform the network design, leading to a novel end-to-end trainable and interpretable deep denoising network, called DeamNet. Note that the unfolding process leads to a promising module called dual element-wise attention mechanism (DEAM) module. To the best of our knowledge, both our ACP constraint and DEAM module have not been reported in the previous literature. Extensive experiments verify the superiority of DeamNet on both synthetic and real noisy image datasets.
****
Topological Planning With Transformers for Vision-and-Language Navigation
Kevin Chen, Junshen K. Chen, Jo Chuang, Marynel Vazquez, Silvio Savarese; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11276-11286
Conventional approaches to vision-and-language navigation (VLN) are trained end-to-end but struggle to perform well in freely traversable environments. Inspired by the robotics community, we propose a modular approach to VLN using topological maps. Given a natural language instruction and topological map, our approach leverages attention mechanisms to predict a navigation plan in the map. The plan is then executed with low-level actions (e.g. forward, rotate) using a robust controller. Experiments show that our method outperforms previous end-to-end approaches, generates interpretable navigation plans, and exhibits intelligent behaviors such as backtracking.
****
FixBi: Bridging Domain Spaces for Unsupervised Domain Adaptation
Jaemin Na, Heechul Jung, Hyung Jin Chang, Wonjun Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1094

-1103
Unsupervised domain adaptation (UDA) methods for learning domain invariant repre sentations have achieved remarkable progress. However, most of the studies were based on direct adaptation from the source domain to the target domain and have suffered from large domain discrepancies. In this paper, we propose a UDA method that effectively handles such large domain discrepancies. We introduce a fixed ratio-based mixup to augment multiple intermediate domains between the source an d target domain. From the augmented-domains, we train the source-dominant model and the target-dominant model that have complementary characteristics. Using our confidence-based learning methodologies, e.g., bidirectional matching with high -confidence predictions and self-penalization using low-confidence predictions, the models can learn from each other or from its own results. Through our propos ed methods, the models gradually transfer domain knowledge from the source to th e target domain. Extensive experiments demonstrate the superiority of our propos ed method on three public benchmarks: Office-31, Office-Home, and VisDA-2017.
****
Generalized Few-Shot Object Detection Without Forgetting
Zhibo Fan, Yuchen Ma, Zeming Li, Jian Sun; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4527-4536
Learning object detection from few examples recently emerged to deal with data-l imited situations. While most previous works merely focus on the performance on few-shot categories, we claim that the ability to detect all classes is crucial as test samples may contain any instances in realistic applications, which requi res the few-shot detector to learn new concepts without forgetting. Through anal ysis on transfer learning based methods, some neglected but beneficial propertie s are utilized to design a simple yet effective few-shot detector, Retentive R-C NN. It consists of Bias-Balanced RPN to debias the pretrained RPN and Re-detecto r to find few-shot class objects without forgetting previous knowledge. Extensiv e experiments on few-shot detection benchmarks show that Retentive R-CNN signifi cantly outperforms state-of-the-art methods on overall performance among all set tings as it can achieve competitive results on few-shot classes and does not deg rade on base class performance at all. Our approach has demonstrated that the lo ng desired never-forgetting learner is available in object detection.
****
Truly Shift-Invariant Convolutional Neural Networks
Anadi Chaman, Ivan Dokmanic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3773-3783
Thanks to the use of convolution and pooling layers, convolutional neural networ ks were for a long time thought to be shift-invariant. However, recent works hav e shown that the output of a CNN can change significantly with small shifts in i nput--a problem caused by the presence of downsampling (stride) layers. The exis ting solutions rely either on data augmentation or on anti-aliasing, both of whi ch have limitations and neither of which enables perfect shift invariance. Addit ionally, the gains obtained from these methods do not extend to image patterns n ot seen during training. To address these challenges, we propose adaptive polyph ase sampling (APS), a simple sub-sampling scheme that allows convolutional neura l networks to achieve 100% consistency in classification performance under shift s, without any loss in accuracy. With APS, the networks exhibit perfect consiste ncy to shifts even before training, making it the first approach that makes conv olutional neural networks truly shift-invariant.
****
Leveraging the Availability of Two Cameras for Illuminant Estimation
Abdelrahman Abdelhamed, Abhijith Punnappurath, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6637-6646
Most modern smartphones are now equipped with two rear-facing cameras -- a main camera for standard imaging and an additional camera to provide wide-angle or te lephoto zoom capabilities. In this paper, we leverage the availability of these two cameras for the task of illumination estimation using a small neural network to perform the illumination prediction. Specifically, if the two cameras' senso

rs have different spectral sensitivities, the two images provide different spect ral measurements of the physical scene. A linear 3x3 color transform that maps b etween these two observations -- and that is unique to a given scene illuminant -- can be used to train a lightweight neural network comprising no more than 146 0 parameters to predict the scene illumination. We demonstrate that this two-cam era approach with a lightweight network provides results on par or better than m uch more complicated illuminant estimation methods operating on a single image. We validate our method's effectiveness through extensive experiments on radiomet ric data, a quasi-real two-camera dataset we generated from an existing single c amera dataset, as well as a new real image dataset that we captured using a smar tphone with two rear-facing cameras.
****
LiDAR-Based Panoptic Segmentation via Dynamic Shifting Network
Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13090-13099
With the rapid advances of autonomous driving, it becomes critical to equip its sensing system with more holistic 3D perception. However, existing works focus o n parsing either the objects (e.g. cars and pedestrians) or scenes (e.g. trees a nd buildings) from the LiDAR sensor. In this work, we address the task of LiDAR-based panoptic segmentation, which aims to parse both objects and scenes in a un ified manner. As one of the first endeavors towards this new challenging task, w e propose the Dynamic Shifting Network (DS-Net), which serves as an effective pa noptic segmentation framework in the point cloud realm. In particular, DS-Net ha s three appealing properties: 1) strong backbone design. DS-Net adopts the cylin der convolution that is specifically designed for LiDAR point clouds. The extrac ted features are shared by the semantic branch and the instance branch which ope rates in a bottom-up clustering style. 2) Dynamic Shifting for complex point dis tributions. We observe that commonly-used clustering algorithms like BFS or DBSC AN are incapable of handling complex autonomous driving scenes with non-uniform point cloud distributions and varying instance sizes. Thus, we present an effici ent learnable clustering module, dynamic shifting, which adapts kernel functions on-the-fly for different instances. 3) Consensus-driven Fusion. Finally, consen sus-driven fusion is used to deal with the disagreement between semantic and ins tance predictions. To comprehensively evaluate the performance of LiDAR-based pa noptic segmentation, we construct and curate benchmarks from two large-scale aut onomous driving LiDAR datasets, SemanticKITTI and nuScenes. Extensive experiment s demonstrate that our proposed DS-Net achieves superior accuracies over current state-of-the-art methods. Notably, we achieve 1st place on the public leaderboa rd of SemanticKITTI, outperforming 2nd place by 2.6% in terms of the PQ metric.
****
Towards Accurate 3D Human Motion Prediction From Incomplete Observations
Qiongjie Cui, Huaijiang Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4801-4810
Predicting accurate and realistic future human poses from historically observed sequences is a fundamental task in the intersection of computer vision, graphics , and artificial intelligence. Recently, continuous efforts have been devoted to addressing this issue, which has achieved remarkable progress. However, the exi sting work is seriously limited by complete observation, that is, once the histo rical motion sequence is incomplete (with missing values), it can only produce u nexpected predictions or even deformities. Furthermore, due to inevitable reason s such as occlusion and the lack of equipment precision, the incompleteness of m otion data occurs frequently, which hinders the practical application of current algorithms. In this work, we first notice this challenging problem, i.e., how t o generate high-fidelity human motion predictions from incomplete observations. To solve it, we propose a novel multi-task graph convolutional network (MT-GCN). Specifically, the model involves two branches, in which the primary task is to focus on forecasting future 3D human actions accurately, while the auxiliary one is to repair the missing value of the incomplete observation. Both of them are integrated into a unified framework to share the spatio-temporal representation,

 which improves the final performance of each collaboratively. On three large-sc
ale datasets, for various data missing scenarios in the real world, extensive ex
periments demonstrate that our approach is consistently superior to the state-of
-the-art methods in which the missing values from incomplete observations are no
t explicitly analyzed.
****
SiamMOT: Siamese Multi-Object Tracking
Bing Shuai, Andrew Berneshawi, Xinyu Li, Davide Modolo, Joseph Tighe; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2021, pp. 12372-12382
In this work, we focus on improving online multi-object tracking (MOT). In parti
cular, we propose a novel region-based Siamese Multi-Object Tracking network, wh
ich we name SiamMOT. SiamMOT is based upon Faster-RCNN and adds a forward tracke
r that models the instance's motion across two frames such that detected instanc
es can be associated in an online fashion. We present two variants of this track
er, an implicit motion model and a novel Siamese-type explicit motion model. We
carry out extensive quantitative experiments on three important MOT datasets: MO
T17, TAO-person and Caltech Roadside Pedestrians, showing the importance of moti
on modelling for MOT and the ability of SiamMOT to substantially outperform the
state-of-the-art. Finally, SiamMOT also outperforms the winners of ACM MM'20 HiE
ve Grand Challenge on the Human in Events dataset. Moreover, SiamMOT is efficien
t, and it runs at 17 FPS for 720P videos on a single modern GPU. We will release
 SiamMOT source code upon acceptance of this paper.
****
Open-Book Video Captioning With Retrieve-Copy-Generate Network
Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, Weiming H
u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2021, pp. 9837-9846
In this paper, we convert traditional video captioning task into a new paradigm,
 i.e., Open-book Video Captioning, which generates natural language under the pr
ompts of video-content-relevant sentences, not limited to the video itself. To a
ddress the open-book video captioning problem, we propose a novel Retrieve-Copy-
Generate network, where a pluggable video-to-text retriever is leveraged to effe
ctively retrieve sentences as hints from the training corpus, and a copy-mechani
sm generator is introduced to dynamically extract expressions from multi-retriev
als. The two modules can be trained end-to-end or separately which is flexible a
nd extensible. Our framework coordinates the conventional retrieval based method
s with orthodox encoder-decoder methods, which can not only draw on the diverse
expressions in the retrieved sentences but also generate natural and accurate co
ntent of the video. Extensive experiments on several benchmark datasets show tha
t our proposed approach performs better than state-of-the-art approaches, indica
ting the effectiveness and promising of the proposed paradigm in the task of vid
eo captioning.
****
MUST-GAN: Multi-Level Statistics Transfer for Self-Driven Person Image Generatio
n
Tianxiang Ma, Bo Peng, Wei Wang, Jing Dong; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13622-13631
Pose-guided person image generation usually involves using paired source-target
images to supervise the training, which significantly increases the data prepara
tion effort and limits the application of the models. To deal with this problem,
 we propose a novel multi-level statistics transfer model, which disentangles an
d transfers multi-level appearance features from person images and merges them w
ith pose features to reconstruct the source person images themselves. So that th
e source images can be used as supervision for self-driven person image generati
on. Specifically, our model extracts multi-level features from the appearance en
coder and learns the optimal appearance representation through attention mechani
sm and attributes statistics. Then we transfer them to a pose-guided generator f
or re-fusion of appearance and pose. Our approach allows for flexible manipulati
on of person appearance and pose properties to perform pose transfer and clothes

style transfer tasks. Experimental results on the DeepFashion dataset demonstrate our method's superiority compared with state-of-the-art supervised and unsupervised methods. In addition, our approach also performs well in the wild.
****
Learning Camera Localization via Dense Scene Matching
Shitao Tang, Chengzhou Tang, Rui Huang, Siyu Zhu, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1831-1841
Camera localization aims to estimate 6 DoF camera poses from RGB images. Traditional methods detect and match interest points between a query image and a pre-built 3D model. Recent learning-based approaches encode scene structures into a specific convolutional neural network(CNN) and thus are able to predict dense coordinates from RGB images. However, most of them require re-training or re-adaption for a new scene and have difficulties in handling large-scale scenes due to limited network capacity. We present a new method for scene agnostic camera localization using dense scene matching (DSM), where the cost volume is constructed between a query image and a scene. The cost volume and the corresponding coordinates are processed by a CNN to predict dense coordinates. Camera poses can then be solved by PnP algorithms. In addition, our method can be extended to temporal domain, giving extra performance boost during testing time. Our scene-agnostic approach achieves comparable accuracy as the existing scene-specific approaches on the 7scenes and Cambridge benchmark. This approach also remarkably outperforms state-of-the-art scene-agnostic dense coordinate regression network SANet.
****
SDD-FIQA: Unsupervised Face Image Quality Assessment With Similarity Distribution Distance
Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, Yuan-Gen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7670-7679
In recent years, Face Image Quality Assessment (FIQA) has become an indispensable part of the face recognition system to guarantee the stability and reliability of recognition performance in an unconstrained scenario. For this purpose, the FIQA method should consider both the intrinsic property and the recognizability of the face image. Most previous works aim to estimate the sample-wise embedding uncertainty or pair-wise similarity as the quality score, which only considers the partial information from the intra-class. However, these methods ignore the valuable information from the inter-class, which is for estimating the recognizability of face image. In this work, we argue that a high-quality face image should be similar to its intra-class samples and dissimilar to its inter-class samples. Thus, we propose a novel unsupervised FIQA method that incorporates Similarity Distribution Distance for Face Image Quality Assessment (SDD-FIQA). Our method generates quality pseudo-labels by calculating the Wasserstein Distance (WD) between the intra-class and inter-class similarity distributions. With these quality pseudo-labels, we are capable of training a regression network for quality prediction. Extensive experiments on benchmark datasets demonstrate that the proposed SDD-FIQA surpasses the state-of-the-arts by an impressive margin. Meanwhile, our method shows good generalization across different recognition systems.
****
Self-Aligned Video Deraining With Transmission-Depth Consistency
Wending Yan, Robby T. Tan, Wenhan Yang, Dengxin Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11966-11976
In this paper, we address the problems of rain streaks and rain accumulation removal in video, by developing a self-aligned network with transmission-depth consistency. Existing video based deraining method focus only on rain streak removal, and commonly use optical flow to align the rain video frames. However, besides rain streaks, rain accummulation can considerably degrade visibility; and, optical flow estimation in a rain video is still erroneous, making the deraining performance tend to be inaccurate. Our method employs deformable convolution layers in our encoder to achieve feature-level frame alignment, and hence avoids using

optical flow. For rain streaks, our method predicts the current frame from its adjacent frames, such that rain streaks that appear randomly in the temporal domain can be removed. For rain accumulation, our method employs transmission-depth consistency to resolve the ambiguity between the depth and water-droplet density. Our network estimates the depth from consecutive rain-accumulation-removal outputs, and we calculate the transmission map using a commonly used physics model. To ensure photometric-temporal and depth-temporal consistencies, our network also estimate the camera poses, so that we can warp one frame to its adjacent frames. Experimental results show that our method is effective in removing both rain streaks and rain accumulation. Our results outperform those of state-of-the-art methods quantitatively and qualitatively.
****

Self-Promoted Prototype Refinement for Few-Shot Class-Incremental Learning
Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6801-6810

Few-shot class-incremental learning is to recognize the new classes given few samples and not forget the old classes. It is a challenging task since representation optimization and prototype reorganization can only be achieved under little supervision. To address this problem, we propose a novel incremental prototype learning scheme. Our scheme consists of a random episode selection strategy that adapts the feature representation to various generated incremental episodes to enhance the corresponding extensibility, and a self-promoted prototype refinement mechanism which strengthens the expression ability of the new class by explicitly considering the dependencies among different classes. Particularly, a dynamic relation projection module is proposed to calculate the relation matrix in a shared embedding space and leverage it as the factor for bootstrapping the update of prototypes. Extensive experiments on three benchmark datasets demonstrate the above-par incremental performance, outperforming state-of-the-art methods by a margin of 13%, 17% and 11%, respectively.
****

PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation
Tal Reiss, Niv Cohen, Liron Bergman, Yedid Hoshen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2806-2814

Anomaly detection methods require high-quality features. In recent years, the anomaly detection community has attempted to obtain better features using advances in deep self-supervised feature learning. Surprisingly, a very promising direction, using pre-trained deep features, has been mostly overlooked. In this paper, we first empirically establish the perhaps expected, but unreported result, that combining pre-trained features with simple anomaly detection and segmentation methods convincingly outperforms, much more complex, state-of-the-art methods. In order to obtain further performance gains in anomaly detection, we adapt pre-trained features to the target distribution. Although transfer learning methods are well established in multi-class classification problems, the one-class classification (OCC) setting is not as well explored. It turns out that naive adaptation methods, which typically work well in supervised learning, often result in catastrophic collapse (feature deterioration) and reduce performance in OCC settings. A popular OCC method, DeepSVDD, advocates using specialized architectures, but this limits the adaptation performance gain. We propose two methods for combating collapse: i) a variant of early stopping that dynamically learns the stopping iteration ii) elastic regularization inspired by continual learning. Our method, PANDA, outperforms the state-of-the-art in the OCC, outlier exposure and anomaly segmentation settings by large margins.
****

Towards Compact CNNs via Collaborative Compression
Yuchao Li, Shaohui Lin, Jianzhuang Liu, Qixiang Ye, Mengdi Wang, Fei Chao, Fan Yang, Jincheng Ma, Qi Tian, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6438-6447

Channel pruning and tensor decomposition have received extensive attention in convolutional neural network compression. However, these two techniques are tradit

ionally deployed in an isolated manner, leading to significant accuracy drop whe n pursuing high compression rates. In this paper, we propose a Collaborative Com pression (CC) scheme, which joints channel pruning and tensor decomposition to c ompress CNN models by simultaneously learning the model sparsity and low-ranknes s. Specifically, we first investigate the compression sensitivity of each layer in the network, and then propose a Global Compression Rate Optimization that tra nsforms the decision problem of compression rate into an optimization problem. A fter that, we propose multi-step heuristic compression to remove redundant compr ession units step-by-step, which fully considers the effect of the remaining com pression space (i.e., unremoved compression units). Our method demonstrates supe rior performance gains over previous ones on various datasets and backbone archi tectures. For example, we achieve 52.9% FLOPs reduction by removing 48.4% parame ters on ResNet-50 with only a Top-1 accuracy drop of 0.56% on ImageNet 2012.
****
Embracing Uncertainty: Decoupling and De-Bias for Robust Temporal Grounding
Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, Chuanping Hu; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 8445-8454
Temporal grounding aims to localize temporal boundaries within untrimmed videos by language queries, but it faces the challenge of two types of inevitable human uncertainties: query uncertainty and label uncertainty. The two uncertainties s tem from human subjectivity, leading to limited generalization ability of tempor al grounding. In this work, we propose a novel DeNet (Decoupling and De-bias) to embrace human uncertainty: Decoupling -- We explicitly disentangle each query i nto a relation feature and a modified feature. The relation feature, which is ma inly based on skeleton-like words (including nouns and verbs), aims to extract b asic and consistent information in the presence of query uncertainty. Meanwhile, modified feature assigned with style-like words (including adjectives, adverbs, etc) represents the subjective information, and thus brings personalized predic tions; De-bias -- We propose a de-bias mechanism to generate diverse predictions , aim to alleviate the bias caused by single-style annotations in the presence o f label uncertainty. Moreover, we put forward new multi-label metrics to diversi fy the performance evaluation. Extensive experiments show that our approach is m ore effective and robust than state-of-the-arts on Charades-STA and ActivityNet Captions datasets.
****
Separating Skills and Concepts for Novel Visual Question Answering
Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, Kate Saenko; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5632-5641
Generalization to out-of-distribution data has been a problem for Visual Questio n Answering (VQA) models. To measure generalization to novel questions, we propo se to separate them into "skills" and "concepts". "Skills" are visual tasks, suc h as counting or attribute recognition, and are applied to "concepts" mentioned in the question, such as objects and people. VQA methods should be able to compo se skills and concepts in novel ways, regardless of whether the specific composi tion has been seen in training, yet we demonstrate that existing models have muc h to improve upon towards handling new compositions. We present a novel method f or learning to compose skills and concepts that separates these two factors impl icitly within a model by learning grounded concept representations and disentang ling the encoding of skills from that of concepts. We enforce these properties w ith a novel contrastive learning procedure that does not rely on external annota tions and can be learned from unlabeled image-question pairs. Experiments demons trate the effectiveness of our approach for improving compositional and groundin g performance.
****
Discrete-Continuous Action Space Policy Gradient-Based Attention for Image-Text Matching
Shiyang Yan, Li Yu, Yuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8096-8105

Image-text matching is an important multi-modal task with massive applications. It tries to match the image and the text with similar semantic information. Existing approaches do not explicitly transform the different modalities into a common space. Meanwhile, the attention mechanism which is widely used in image-text matching models does not have supervision. We propose a novel attention scheme which projects the image and text embedding into a common space and optimises the attention weights directly towards the evaluation metrics. The proposed attention scheme can be considered as a kind of supervised attention and requiring no additional annotations. It is trained via a novel Discrete-continuous action space policy gradient algorithm, which is more effective in modelling complex action space than previous continuous action space policy gradient. We evaluate the proposed methods on two widely-used benchmark datasets: Flickr30k and MS-COCO, outperforming the previous approaches by a large margin.
****
Scalable Differential Privacy With Sparse Network Finetuning
Zelun Luo, Daniel J. Wu, Ehsan Adeli, Li Fei-Fei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5059-5068
We propose a novel method for privacy-preserving training of deep neural networks leveraging public, out-domain data. While differential privacy (DP) has emerged as a mechanism to protect sensitive data in training datasets, its application to complex visual recognition tasks remains challenging. Traditional DP methods, such as Differentially-Private Stochastic Gradient Descent (DP-SGD), only perform well on simple datasets and shallow networks, while recent transfer learning-based DP methods often make unrealistic assumptions about the availability and distribution of public data. In this work, we argue that minimizing the number of trainable parameters is the key to improving the privacy-performance tradeoff of DP on complex visual recognition tasks. We also propose a novel transfer learning paradigm that finetunes a very sparse subnetwork with DP, inspired by this argument. We conduct extensive experiments and ablation studies on two visual recognition tasks: CIFAR-100 -> CIFAR-10 (standard DP setting) and the CD-FSL challenge (few-shot, multiple levels of domain shifts) and demonstrate competitive experimental performance.
****
Video Object Segmentation Using Global and Instance Embedding Learning
Wenbin Ge, Xiankai Lu, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16836-16845
In this paper, we propose a feature embedding based video object segmentation (VOS) method which is simple, fast and effective. The current VOS task involves two main challenges: object instance differentiation and cross-frame instance alignment. Most state-of-the-art matching based VOS methods simplify this task into a binary segmentation task and tackle each instance independently. In contrast, we decompose the VOS task into two subtasks: global embedding learning that segments foreground objects of each frame in a pixel-to-pixel manner, and instance feature embedding learning that separates instances. The outputs of these two subtasks are fused to obtain the final instance masks quickly and accurately. Through using the relation among different instances per-frame as well as temporal relation across different frames, the proposed network learns to differentiate multiple instances and associate them properly in one feed-forward manner. Extensive experimental results on the challenging DAVIS and Youtube-VOS datasets show that our method achieves better performances than most counterparts in each case.
****
Scene Text Retrieval via Joint Text Detection and Similarity Learning
Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4558-4567
Scene text retrieval aims to localize and search all text instances from an image gallery, which are the same or similar with a given query text. Such a task is usually realized by matching a query text to the recognized words, outputted by an end-to-end scene text spotter. In this paper, we address this problem by directly learning a cross-modal similarity between a query text and each text insta

nce from natural images. Specifically, we establish an end-to-end trainable netw
ork, jointly optimizing the procedures of scene text detection and cross-modal s
imilarity learning. In this way, scene text retrieval can be simply performed by
 ranking the detected text instances with the learned similarity. Experiments on
 three benchmark datasets demonstrate our method consistently outperforms the st
ate-of-the-art scene text spotting/retrieval approaches. In particular, the prop
osed framework of joint detection and similarity learning achieves significantly
 better performance than separated methods.
****
Learning Continuous Image Representation With Local Implicit Image Function
Yinbo Chen, Sifei Liu, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8628-8638
How to represent an image? While the visual world is presented in a continuous m
anner, machines store and see the images in a discrete way with 2D arrays of pix
els. In this paper, we seek to learn a continuous representation for images. Ins
pired by the recent progress in 3D reconstruction with implicit neural represent
ation, we propose Local Implicit Image Function (LIIF), which takes an image coo
rdinate and the 2D deep features around the coordinate as inputs, predicts the R
GB value at a given coordinate as an output. Since the coordinates are continuou
s, LIIF can be presented in arbitrary resolution. To generate the continuous rep
resentation for images, we train an encoder with LIIF representation via a self-
supervised task with super-resolution. The learned continuous representation can
 be presented in arbitrary resolution even extrapolate to x30 higher resolution,
 where the training tasks are not provided. We further show that LIIF representa
tion builds a bridge between discrete and continuous representation in 2D, it na
turally supports the learning tasks with size-varied image ground-truths and sig
nificantly outperforms the method with resizing the ground-truths.
****
Locally Aware Piecewise Transformation Fields for 3D Human Mesh Registration
Shaofei Wang, Andreas Geiger, Siyu Tang; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7639-7648
Registering point clouds of dressed humans to parametric human models is a chall
enging task in computer vision. Traditional approaches often rely on heavily eng
ineered pipelines that require accurate manual initialization of human poses and
 tedious post-processing. More recently, learning-based methods are proposed in
hope to automate this process. We observe that pose initialization is key to acc
urate registration but existing methods often fail to provide accurate pose init
ialization. One major obstacle is that, despite recent effort on rotation repres
entation learning in neural networks, regressing joint rotations from point clou
ds or images of humans is still very challenging. To this end, we propose novel
piecewise transformation fields (PTF), a set of functions that learn 3D translat
ion vectors to map any query point in posed space to its correspond position in
rest-pose space. We combine PTF with multi-class occupancy networks, obtaining a
 novel learning-based framework that learns to simultaneously predict shape and
per-point correspondences between the posed space and the canonical space for cl
othed human. Our key insight is that the translation vector for each query point
 can be effectively estimated using the point-aligned local features; consequent
ly, rigid per bone transformations and joint rotations can be obtained efficient
ly via a least-square fitting given the estimated point correspondences, circumv
enting the challenging task of directly regressing joint rotations from neural n
etworks. Furthermore, the proposed PTF facilitate canonicalized occupancy estima
tion, which greatly improves generalization capability and result in more accura
te surface reconstruction with only half of the parameters compared with the sta
te-of-the-art. Both qualitative and quantitative studies show that fitting param
etric models with poses initialized by our network results in much better regist
ration quality, especially for extreme poses.
****
Graph Attention Tracking
Dongyan Guo, Yanyan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, Chunhua Shen; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (

CVPR), 2021, pp. 9543-9552
Siamese network based trackers formulate the visual tracking task as a similarit y matching problem. Almost all popular Siamese trackers realize the similarity l earning via convolutional feature cross-correlation between a target branch and a search branch. However, since the size of target feature region needs to be pr e-fixed, these cross-correlation base methods suffer from either reserving much adverse background information or missing a great deal of foreground information . Moreover, the global matching between the target and search region also largel y neglects the target structure and part-level information. In this paper, to so lve the above issues, we propose a simple target-aware Siamese graph attention n etwork for general object tracking. We propose to establish part-to-part corresp ondence between the target and the search region with a complete bipartite graph , and apply the graph attention mechanism to propagate target information from t he template feature to the search feature. Further, instead of using the pre-fix ed region cropping for template-feature-area selection, we investigate a target-aware area selection mechanism to fit the size and aspect ratio variations of di fferent objects. Experiments on challenging benchmarks including GOT-10k, UAV123 , OTB-100 and LaSOT demonstrate that the proposed SiamGAT outperforms many state -of-the-art trackers and achieves leading performance. Code is available at: htt ps://git.io/SiamGAT
****
ReDet: A Rotation-Equivariant Detector for Aerial Object Detection
Jiaming Han, Jian Ding, Nan Xue, Gui-Song Xia; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2786-2795
Recently, object detection in aerial images has gained much attention in compute r vision. Different from objects in natural images, aerial objects are often dis tributed with arbitrary orientation. Therefore, the detector requires more param eters to encode the orientation information, which are often highly redundant an d inefficient. Moreover, as ordinary CNNs do not explicitly model the orientatio n variation, large amounts of rotation augmented data is needed to train an accu rate object detector. In this paper, we propose a Rotation-equivariant Detector (ReDet) to address these issues, which explicitly encodes rotation equivariance and rotation invariance. More precisely, we incorporate rotation-equivariant net works into the detector to extract rotation-equivariant features, which can accu rately predict the orientation and lead to a huge reduction of model size. Based on the rotation-equivariant features, we also present Rotation-invariant RoI Al ign (RiRoI Align), which adaptively extracts rotation-invariant features from eq uivariant features according to the orientation of RoI. Extensive experiments on several challenging aerial image datasets DOTA-v1.0, DOTA-v1.5 and HRSC2016, sh ow that our method can achieve state-of-the-art performance on the task of aeria l object detection. Compared with previous best results, our ReDet gains 1.2, 3. 5 and 2.6 mAP on DOTA-v1.0, DOTA-v1.5 and HRSC2016 respectively while reducing t he number of parameters by 60% (313 Mb vs. 121 Mb). The code is available at: ht tps://github.com/csuhan/ReDet.
****
Action Shuffle Alternating Learning for Unsupervised Action Segmentation
Jun Li, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2021, pp. 12628-12636
This paper addresses unsupervised action segmentation. Prior work captures the f rame-level temporal structure of videos by a feature embedding that encodes time locations of frames in the video. We advance prior work with a new self-supervi sed learning (SSL) of a feature embedding that accounts for both frame- and acti on-level structure of videos. Our SSL trains an RNN to recognize positive and ne gative action sequences, and the RNN's hidden layer is taken as our new action-l evel feature embedding. The positive and negative sequences consist of action se gments sampled from videos, where in the former the sampled action segments resp ect their time ordering in the video, and in the latter they are shuffled. As su pervision of actions is not available and our SSL requires access to action segm ents, we specify an HMM that explicitly models action lengths, and infer a MAP a ction segmentation with the Viterbi algorithm. The resulting action segmentation

 is used as pseudo-ground truth for estimating our action-level feature embeddin g and updating the HMM. We alternate the above steps within the Generalized EM f ramework, which ensures convergence. Our evaluation on the Breakfast, YouTube In structions, and 50Salads datasets gives superior results to those of the state o f the art.
****
Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences
Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2554-2562
Human multimodal emotion recognition involves time-series data of different moda lities, such as natural language, visual motions, and acoustic behaviors. Due to the variable sampling rates for sequences from different modalities, the collec ted multimodal streams are usually unaligned. The asynchrony across modalities i ncreases the difficulty on conducting efficient multimodal fusion. Hence, this w ork mainly focuses on multimodal fusion from unaligned multimodal sequences. To this end, we propose the Progressive Modality Reinforcement (PMR) approach based on the recent advances of crossmodal transformer. Our approach introduces a mes sage hub to exchange information with each modality. The message hub sends commo n messages to each modality and reinforces their features via crossmodal attenti on. In turn, it also collects the reinforced features from each modality and use s them to generate a reinforced common message. By repeating the cycle process, the common message and the modalities' features can progressively complement eac h other. Finally, the reinforced features are used to make predictions for human emotion. Comprehensive experiments on different human multimodal emotion recogn ition benchmarks clearly demonstrate the superiority of our approach.
****
OpenMix: Reviving Known Knowledge for Discovering Novel Visual Categories in an Open World
Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 021, pp. 9462-9470
In this paper, we tackle the problem of discovering new classes in unlabeled vis ual data given labeled data from disjoint classes. Existing methods typically fi rst pre-train a model with labeled data, and then identify new classes in unlabe led data via unsupervised clustering. However, the labeled data that provide ess ential knowledge are often underexplored in the second step. The challenge is th at the labeled and unlabeled examples are from non-overlapping classes, which ma kes it difficult to build a learning relationship between them. In this work, we introduce OpenMix to mix the unlabeled examples from an open set and the labele d examples from known classes, where their non-overlapping labels and pseudo-lab els are simultaneously mixed into a joint label distribution. OpenMix dynamicall y compounds examples in two ways. First, we produce mixed training images by inc orporating labeled examples with unlabeled examples. With the benefit of unique prior knowledge in novel class discovery, the generated pseudo-labels will be mo re credible than the original unlabeled predictions. As a result, OpenMix helps preventing the model from overfitting on unlabeled samples that may be assigned with wrong pseudo-labels. Second, the first way encourages the unlabeled example s with high class-probabilities to have considerable accuracy. We introduce thes e examples as reliable anchors and further integrate them with unlabeled samples . This enables us to generate more combinations in unlabeled examples and exploi t finer object relations among the new classes. Experiments on three classificat ion datasets demonstrate the effectiveness of the proposed OpenMix, which is sup erior to state-of-the-art methods in novel class discovery.
****
Combining Semantic Guidance and Deep Reinforcement Learning for Generating Human Level Paintings
Jaskirat Singh, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16387-16396

Generation of stroke-based non-photorealistic imagery, is an important problem i
n the computer vision community. As an endeavor in this direction, substantial r
ecent research efforts have been focused on teaching machines "how to paint", in
 a manner similar to a human painter. However, the applicability of previous met
hods has been limited to datasets with little variation in position, scale and s
aliency of the foreground object. As a consequence, we find that these methods s
truggle to cover the granularity and diversity possessed by real world images. T
o this end, we propose a Semantic Guidance pipeline with 1) a bi-level painting
procedure for learning the distinction between foreground and background brush s
trokes at training time. 2) We also introduce invariance to the position and sca
le of the foreground object through a neural alignment model, which combines obj
ect localization and spatial transformer networks in an end to end manner, to zo
om into a particular semantic instance. 3) The distinguishing features of the in
-focus object are then amplified by maximizing a novel guided backpropagation ba
sed focus reward. The proposed agent does not require any supervision on human s
troke-data and successfully handles variations in foreground object attributes,
thus, producing much higher quality canvases for the CUB-200 Birds and Stanford
Cars-196 datasets. Finally, we demonstrate the further efficacy of our method on
 complex datasets with multiple foreground object instances by evaluating an ext
ension of our method on the challenging Virtual-KITTI dataset. Source code and m
odels are available at https://github.com/1jsingh/semantic-guidance.
****
Event-Based Bispectral Photometry Using Temporally Modulated Illumination
Tsuyoshi Takatani, Yuzuha Ito, Ayaka Ebisu, Yinqiang Zheng, Takahito Aoto; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2021, pp. 15638-15647
Analysis of bispectral difference plays a critical role in various applications
that involve rays propagating in a light absorbing medium. In general, the bispe
ctral difference is obtained by subtracting signals at two individual wavelength
s captured by ordinary digital cameras, which tends to inherit the drawbacks of
conventional cameras in dynamic range, response speed and quantization precision
. In this paper, we propose a novel method to obtain a bispectral difference ima
ge using an event camera with temporally modulated illumination. Our method is r
ooted in a key observation on the analogy between the bispectral photometry prin
ciple of the participating medium and the event generating mechanism in an event
 camera. By carefully modulating the bispectral illumination, our method allows
to read out the bispectral difference directly from triggered events. Experiment
s using a prototype imaging system have verified the feasibility of this novel u
sage of event cameras in photometry based vision tasks, such as 3D shape reconst
ruction in water.
****
LiDAR-Aug: A General Rendering-Based Augmentation Framework for 3D Object Detect
ion
Jin Fang, Xinxin Zuo, Dingfu Zhou, Shengze Jin, Sen Wang, Liangjun Zhang; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2021, pp. 4710-4720
Annotating the LiDAR point cloud is crucial for deep learning-based 3D object de
tection tasks. Due to expensive labeling costs, data augmentation has been taken
 as a necessary module and plays an important role in training the neural networ
k. "Copy" and "paste" (i.e., GT-Aug) is the most commonly used data augmentation
 strategy, however, the occlusion between objects has not been taken into consid
eration. To handle the above limitation, we propose a rendering-based LiDAR augm
entation framework (i.e., LiDAR-Aug) to enrich the training data and boost the p
erformance of LiDAR-based 3D object detectors. The proposed LiDAR-Aug is a plug-
and-play module that can be easily integrated into different types of 3D object
detection frameworks. Compared to the traditional object augmentation methods, L
iDAR-Aug is more realistic and effective. Finally, we verify the proposed framew
ork on the public KITTI dataset with different 3D object detectors. The experime
ntal results show the superiority of our method compared to other data augmentat
ion strategies. We plan to make our data and code public to help other researche

rs reproduce our results.
****
Semantic-Aware Knowledge Distillation for Few-Shot Class-Incremental Learning
Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, Mehrtash Harandi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2534-2543
Few-shot class incremental learning (FSCIL) portrays the problem of learning new concepts gradually, where only a few examples per concept are available to the learner. Due to the limited number of examples for training, the techniques developed for standard incremental learning cannot be applied verbatim to FSCIL. In this work, we introduce a distillation algorithm to address the problem of FSCIL and propose to make use of semantic information during training. To this end, we make use of word embeddings as semantic information which is cheap to obtain and which facilitate the distillation process. Furthermore, we propose a method based on an attention mechanism on multiple parallel embeddings of visual data to align visual and semantic vectors, which reduces issues related to catastrophic forgetting. Via experiments on MiniImageNet, CUB200, and CIFAR100 dataset, we establish new state-of-the-art results by outperforming existing approaches.
****
General Instance Distillation for Object Detection
Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, Erjin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7842-7851
In recent years, knowledge distillation has been proved to be an effective solution for model compression. This approach can make lightweight student models acquire the knowledge extracted from cumbersome teacher models. However, previous distillation methods of detection have weak generalization for different detection frameworks and rely heavily on ground truth (GT), ignoring the valuable relation information between instances. Thus, we propose a novel distillation method for detection tasks based on discriminative instances without considering the positive or negative distinguished by GT, which is called general instance distillation (GID). Our approach contains a general instance selection module (GISM) to make full use of feature-based, relation-based and response-based knowledge for distillation. Extensive results demonstrate that the student model achieves significant AP improvement and even outperforms the teacher in various detection frameworks. Specifically, RetinaNet with ResNet-50 achieves 39.1% in mAP with GID on COCO dataset, which surpasses the baseline 36.2% by 2.9%, and even better than the ResNet-101 based teacher model with 38.1% AP.
****
Joint Noise-Tolerant Learning and Meta Camera Shift Adaptation for Unsupervised Person Re-Identification
Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4855-4864
This paper considers the problem of unsupervised person re-identification (re-ID), which aims to learn discriminative models with unlabeled data. One popular method is to obtain pseudo-label by clustering and use them to optimize the model. Although this kind of approach has shown promising accuracy, it is hampered by 1) noisy labels produced by clustering and 2) feature variations caused by camera shift. The former will lead to incorrect optimization and thus hinders the model accuracy. The latter will result in assigning the intra-class samples of different cameras to different pseudo-label, making the model sensitive to camera variations. In this paper, we propose a unified framework to solve both problems. Concretely, we propose a Dynamic and Symmetric Cross-Entropy loss (DSCE) to deal with noisy samples and a camera-aware meta-learning algorithm (MetaCam) to adapt camera shift. DSCE can alleviate the negative effects of noisy samples and accommodate the change of clusters after each clustering step. MetaCam simulates cross-camera constraint by splitting the training data into meta-train and meta-test based on camera IDs. With the interacted gradient from meta-train and meta-test, the model is enforced to learn camera-invariant features. Extensive experime

nts on three re-ID benchmarks show the effectiveness and the complementary of th
e proposed DSCE and MetaCam. Our method outperforms the state-of-the-art methods
 on both fully unsupervised re-ID and unsupervised domain adaptive re-ID.
****
Mutual Graph Learning for Camouflaged Object Detection
Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, Deng-Ping Fan; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2021, pp. 12997-13007
Automatically detecting/segmenting object(s) that blend in with their surroundin
gs is difficult for current models. A major challenge is that the intrinsic simi
larities between such foreground objects and background surroundings make the fe
atures extracted by deep model indistinguishable. To overcome this challenge, an
 ideal model should be able to seek valuable, extra clues from the given scene a
nd incorporate them into a joint learning framework for representation co-enhanc
ement. With this inspiration, we design a novel Mutual Graph Learning (MGL) mode
l, which generalizes the idea of conventional mutual learning from regular grids
 to the graph domain. Specifically, MGL decouples an image into two task-specifi
c feature maps -- one for roughly locating the target and the other for accurate
ly capturing its boundary details -- and fully exploits the mutual benefits by r
ecurrently reasoning their high-order relations through graphs. Importantly, in
contrast to most mutual learning approaches that use a shared function to model
all between-task interactions, MGL is equipped with typed functions for handling
 different complementary relations to maximize information interactions. Experim
ents on challenging datasets, including CHAMELEON, CAMO and COD10K, demonstrate
the effectiveness of our MGL with superior performance to existing state-of-the-
art methods.
****
Single Pair Cross-Modality Super Resolution
Guy Shacht, Dov Danon, Sharon Fogel, Daniel Cohen-Or; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6378-6
387
Non-visual imaging sensors are widely used in the industry for different purpose
s. Those sensors are more expensive than visual (RGB) sensors, and usually produ
ce images with lower resolution. To this end, Cross-Modality Super-Resolution me
thods were introduced, where an RGB image of a high-resolution assists in increa
sing the resolution of a low-resolution modality. However, fusing images from di
fferent modalities is not a trivial task, since each multi-modal pair varies gre
atly in its internal correlations. For this reason, traditional state-of-the-art
s which are trained on external datasets often struggle with yielding an artifac
t-free result that is still loyal to the target modality characteristics. We pre
sent CMSR, a single-pair approach for Cross-Modality Super-Resolution. The netwo
rk is internally trained on the two input images only, in a self-supervised mann
er, learns their internal statistics and correlations, and applies them to upsam
ple the target modality. CMSR contains an internal transformer which is trained
on-the-fly together with the up-sampling process itself and without supervision,
 to allow dealing with pairs that are only weakly aligned. We show that CMSR pro
duces state-of-the-art super resolved images, yet without introducing artifacts
or irrelevant details that originate from the RGB image only.
****
Target-Aware Object Discovery and Association for Unsupervised Video Multi-Objec
t Segmentation
Tianfei Zhou, Jianwu Li, Xueyi Li, Ling Shao; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6985-6994
This paper addresses the task of unsupervised video multi-object segmentation. C
urrent approaches follow a two-stage paradigm: 1) detect object proposals using
pre-trained Mask R-CNN, and 2) conduct generic feature matching for temporal ass
ociation using re-identification techniques. However, the generic features, wide
ly used in both stages, are not reliable for characterizing unseen objects, lead
ing to poor generalization. To address this, we introduce a novel approach for m
ore accurate and efficient spatio-temporal segmentation. In particular, to addre

ss instance discrimination, we propose to combine foreground region estimation a
nd instance grouping together in one network, and additionally introduce tempora
l guidance for segmenting each frame, enabling more accurate object discovery. F
or temporal association, we complement current video object segmentation archite
ctures with a discriminative appearance model, capable of capturing more fine-gr
ained target-specific information. Given object proposals from the instance disc
rimination network, three essential strategies are adopted to achieve accurate s
egmentation: 1) target-specific tracking using a memory-augmented appearance mod
el; 2) target-agnostic verification to trace possible tracklets for the proposal
; 3) adaptive memory updating using the verified segments. We evaluate the propo
sed approach on DAVIS_ 17 and YouTube-VIS, and the results demonstrate that it
outperforms state-of-the-art methods both in segmentation accuracy and inference
 speed.
****
Cross-View Regularization for Domain Adaptive Panoptic Segmentation
Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10133-101
44
Panoptic segmentation unifies semantic segmentation and instance segmentation wh
ich has been attracting increasing attention in recent years. On the other hand,
 most existing research was conducted under a supervised learning setup whereas
domain adaptive panoptic segmentation which is critical in different tasks and a
pplications is largely neglected. We design a domain adaptive panoptic segmentat
ion network that exploits inter-style consistency and inter-task regularization
for optimal domain adaptive panoptic segmentation. The inter-style consistency l
everages geometric invariance across the same image of the different styles whic
h ` fabricates' certain self-supervisions to guide the network to learn domain-i
nvariant features. The inter-task regularization exploits the complementary natu
re of instance segmentation and semantic segmentation and uses it as a constrain
t for better feature alignment across domains. Extensive experiments over multip
le domain adaptive panoptic segmentation tasks (e.g. synthetic-to-real and real-
to-real) show that our proposed network achieves superior segmentation performan
ce as compared with the state-of-the-art.
****
End-to-End Learning for Joint Image Demosaicing, Denoising and Super-Resolution
Wenzhu Xing, Karen Egiazarian; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2021, pp. 3507-3516
Image denoising, demosaicing and super-resolution are key problems of image rest
oration well studied in the recent decades. Often, in practice, one has to solve
 these problems simultaneously. A problem of finding a joint solution of the mul
tiple image restoration tasks just begun to attract an increased attention of re
searchers. In this paper, we propose an end-to-end solution for the joint demosa
icing, denoising and super-resolution based on a specially designed deep convolu
tional neural network (CNN). We systematically study different methods to solve
this problem and compared them with the proposed method. Extensive experiments c
arried out on large image datasets demonstrate that our method outperforms the s
tate-of-the-art both quantitatively and qualitatively. Finally, we have applied
various loss functions in the proposed scheme and demonstrate that by using the
mean absolute error as a loss function, we can obtain superior results in compar
ison to other cases.
****
Keep Your Eyes on the Lane: Real-Time Attention-Guided Lane Detection
Lucas Tabelini, Rodrigo Berriel, Thiago M. Paixao, Claudine Badue, Alberto F. De
 Souza, Thiago Oliveira-Santos; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 294-302
Modern lane detection methods have achieved remarkable performances in complex r
eal-world scenarios, but many have issues maintaining real-time efficiency, whic
h is important for autonomous vehicles. In this work, we propose LaneATT: an anc
hor-based deep lane detection model, which, akin to other generic deep object de
tectors, uses the anchors for the feature pooling step. Since lanes follow a reg

ular pattern and are highly correlated, we hypothesize that in some cases global information may be crucial to infer their positions, especially in conditions such as occlusion, missing lane markers, and others. Thus, this work proposes a novel anchor-based attention mechanism that aggregates global information. The model was evaluated extensively on three of the most widely used datasets in the literature. The results show that our method outperforms the current state-of-the-art methods showing both higher efficacy and efficiency. Moreover, an ablation study is performed along with a discussion on efficiency trade-off options that are useful in practice. Code and models are available at https://github.com/lucastabelini/LaneATT.
****
Lesion-Aware Transformers for Diabetic Retinopathy Grading
Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10938-10947
Diabetic retinopathy (DR) is the leading cause of permanent blindness in the working-age population. And automatic DR diagnosis can assist ophthalmologists to design tailored treatments for patients, including DR grading and lesion discovery. However, most of existing methods treat DR grading and lesion discovery as two independent tasks, which require lesion annotations as a learning guidance and limits the actual deployment. To alleviate this problem, we propose a novel lesion-aware transformer (LAT) for DR grading and lesion discovery jointly in a unified deep model via an encoder-decoder structure including a pixel relation based encoder and a lesion filter based decoder. The proposed LAT enjoys several merits. First, to the best of our knowledge, this is the first work to formulate lesion discovery as a weakly supervised lesion localization problem via a transformer decoder. Second, to learn lesion filters well with only image-level labels, we design two effective mechanisms including lesion region importance and lesion region diversity for identifying diverse lesion regions. Extensive experimental results on three challenging benchmarks including Messidor-1, Messidor-2 and EyePACS demonstrate that the proposed LAT performs favorably against state-of-the-art DR grading and lesion discovery methods.
****
Involution: Inverting the Inherence of Convolution for Visual Recognition
Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12321-12330
Convolution has been the core ingredient of modern neural networks, triggering the surge of deep learning in vision. In this work, we rethink the inherent principles of standard convolution for vision tasks, specifically spatial-agnostic and channel-specific. Instead, we present a novel atomic operation for deep neural networks by inverting the aforementioned design principles of convolution, coined as involution. We additionally demystify the recent popular self-attention operator and subsume it into our involution family as an over-complicated instantiation. The proposed involution operator could be leveraged as fundamental bricks to build the new generation of neural networks for visual recognition, powering different deep learning models on several prevalent benchmarks, including ImageNet classification, COCO detection and segmentation, together with Cityscapes segmentation. Our involution-based models improve the performance of convolutional baselines using ResNet-50 by up to 1.6% top-1 accuracy, 2.5% and 2.4% bounding box AP, and 4.7% mean IoU absolutely while compressing the computational cost to 66%, 65%, 72%, and 57% on the above benchmarks, respectively. Code and pre-trained models for all the tasks are available at https://github.com/d-li14/involution.
****
QPIC: Query-Based Pairwise Human-Object Interaction Detection With Image-Wide Contextual Information
Masato Tamura, Hiroki Ohashi, Tomoaki Yoshinaga; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10410-10419
We propose a simple, intuitive yet powerful method for human-object interaction

(HOI) detection. HOIs are so diverse in spatial distribution in an image that existing CNN-based methods face the following three major drawbacks; they cannot leverage image-wide features due to CNN's locality, they rely on a manually defined location-of-interest for the feature aggregation, which sometimes does not cover contextually important regions, and they cannot help but mix up the features for multiple HOI instances if they are located closely. To overcome these drawbacks, we propose a transformer-based feature extractor, in which an attention mechanism and query-based detection play key roles. The attention mechanism is effective in aggregating contextually important information image-wide, while the queries, which we design in such a way that each query captures at most one human-object pair, can avoid mixing up the features from multiple instances. This transformer-based feature extractor produces so effective embeddings that the subsequent detection heads may be fairly simple and intuitive. The extensive analysis reveals that the proposed method successfully extracts contextually important features, and thus outperforms existing methods by large margins (5.37 mAP on HICO-DET, and 5.6 mAP on V-COCO). The source codes are available at https://github.com/hitachi-rd-cv/qpic.
****
Home Action Genome: Cooperative Compositional Action Understanding
Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11184-11193
Existing research on action recognition treats activities as monolithic events occurring in videos. Recently, the benefits of formulating actions as a combination of atomic-actions have shown promise in improving action understanding with the emergence of datasets containing such annotations, allowing us to learn representations capturing this information. However, there remains a lack of studies that extend action composition and leverage multiple viewpoints and multiple modalities of data for representation learning. To promote research in this direction, we introduce Home Action Genome (HOMAGE): a multi-view action dataset with multiple modalities and view-points supplemented with hierarchical activity and atomic action labels together with dense scene composition labels. Leveraging rich multi-modal and multi-view settings, we propose Cooperative Compositional Action Understanding (CCAU), a cooperative learning framework for hierarchical action recognition that is aware of compositional action elements. CCAU shows consistent performance improvements across all modalities. Furthermore, we demonstrate the utility of co-learning compositions in few-shot action recognition by achieving 28.6% mAP with just a single sample.
****
Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies
Jinzheng Cai, Youbao Tang, Ke Yan, Adam P. Harrison, Jing Xiao, Gigin Lin, Le Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15159-15169
Monitoring treatment response in longitudinal studies plays an important role in clinical practice. Accurately identifying lesions across serial imaging follow-up is the core to the monitoring procedure. Typically this incorporates both image and anatomical considerations. However, matching lesions manually is labor-intensive and time-consuming. In this work, we present deep lesion tracker (DLT), a deep learning approach that uses both appearance- and anatomical-based signals. To incorporate anatomical constraints, we propose an anatomical signal encoder, which prevents lesions being matched with visually similar but spurious regions. In addition, we present a new formulation for Siamese networks that avoids the heavy computational loads of 3D cross-correlation. To present our network with greater varieties of images, we also propose a self-supervised learning strategy to train trackers with unpaired images, overcoming barriers to data collection. To train and evaluate our tracker, we introduce and release the first lesion tracking benchmark, consisting of 3891 lesion pairs from the public DeepLesion database. The proposed method, DLT, locates lesion centers with a mean error distance of 7mm. This is 5% better than a leading registration algorithm while running 14 times faster with whole CT volumes. We demonstrate even greater improvement

s over detector or similarity-learning alternatives. DLT also generalizes well on an external clinical test set of 100% longitudinal studies, achieving 88% accuracy. Finally, we plug DLT into an automatic tumor monitoring workflow where it leads to an accuracy of 85% in assessing lesion treatment responses, which is only 0.46% lower than the accuracy of manual inputs.
****

Learning To Warp for Style Transfer

Xiao-Chang Liu, Yong-Liang Yang, Peter Hall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3702-3711

Since its inception in 2015, Style Transfer has focused on texturing a content image using an art exemplar. Recently, the geometric changes that artists make have been acknowledged as an important component of style. Our contribution is to propose a neural network that, uniquely, learns a mapping from a 4D array of inter-feature distances to a non-parametric 2D warp field. The system is generic in not being limited by semantic class, a single learned model will suffice; all examples in this paper are output from one model. Our approach combines the benefits of the high speed of Liu et al. with the non-parametric warping of Kim et al. Furthermore, our system extends the normal NST paradigm: although it can be used with a single exemplar, we also allow two style exemplars: one for texture and another for geometry. This supports far greater flexibility in use cases than single exemplars can provide.
****

Towards Extremely Compact RNNs for Video Recognition With Fully Decomposed Hierarchical Tucker Structure

Miao Yin, Siyu Liao, Xiao-Yang Liu, Xiaodong Wang, Bo Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12085-12094

Recurrent Neural Networks (RNNs) have been widely used in sequence analysis and modeling. However, when processing high-dimensional data, RNNs typically require very large model sizes, thereby bringing a series of deployment challenges. Although various prior works have been proposed to reduce the RNN model sizes, executing RNN models in resource-restricted environments is still a very challenging problem. In this paper, we propose to develop extremely compact RNN models with fully decomposed hierarchical Tucker (FDHT) structure. The HT decomposition does not only provide much higher storage cost reduction than the other tensor decomposition approaches but also brings better accuracy performance improvement for the compact RNN models. Meanwhile, unlike the existing tensor decomposition-based methods that can only decompose the input-to-hidden layer of RNNs, our proposed fully decomposition approach enables the comprehensive compression for the entire RNN models with maintaining very high accuracy. Our experimental results on several popular video recognition datasets show that our proposed fully decomposed hierarchical tucker-based LSTM (FDHT-LSTM) is extremely compact and highly efficient. To the best of our knowledge, FDHT-LSTM, for the first time, consistently achieves very high accuracy with only few thousand parameters (3,132 to 8,808) on different datasets. Compared with the state-of-the-art compressed RNN models, such as TT-LSTM, TR-LSTM and BT-LSTM, our FDHT-LSTM simultaneously enjoys both order-of-magnitude (3,985x to 10,711x) fewer parameters and significant accuracy improvement (0.6% to 12.7%).
****

Self-Supervised Multi-Frame Monocular Scene Flow

Junhwa Hur, Stefan Roth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2684-2694

Estimating 3D scene flow from a sequence of monocular images has been gaining increased attention due to the simple, economical capture setup. Owing to the severe ill-posedness of the problem, the accuracy of current methods has been limited, especially that of efficient, real-time approaches. In this paper, we introduce a multi-frame monocular scene flow network based on self-supervised learning, improving the accuracy over previous networks while retaining real-time efficiency. Based on an advanced two-frame baseline with a split-decoder design, we propose (i) a multi-frame model using a triple frame input and convolutional LSTM c

onnections, (ii) an occlusion-aware census loss for better accuracy, and (iii) a gradient detaching strategy to improve training stability. On the KITTI dataset, we observe state-of-the-art accuracy among monocular scene flow methods based on self-supervised learning.
****
Enriching ImageNet With Human Similarity Judgments and Psychological Embeddings
Brett D. Roads, Bradley C. Love; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3547-3557
Advances in supervised learning approaches to object recognition flourished in part because of the availability of high-quality datasets and associated benchmarks. However, these benchmarks---such as ILSVRC---are relatively task-specific, focusing predominately on predicting class labels. We introduce a publicly-available dataset that embodies the task-general capabilities of human perception and reasoning. The Human Similarity Judgments extension to ImageNet (ImageNet-HSJ) is composed of a large set of human similarity judgments that supplements the existing ILSVRC validation set. The new dataset supports a range of task and performance metrics, including evaluation of unsupervised algorithms. We demonstrate two methods of assessment: using the similarity judgments directly and using a psychological embedding trained on the similarity judgments. This embedding space contains an order of magnitude more points (i.e., images) than previous efforts based on human judgments. We were able to scale to the full 50,000 image ILSVRC validation set through a selective sampling process that used variational Bayesian inference and model ensembles to sample aspects of the embedding space that were most uncertain. To demonstrate the utility of ImageNet-HSJ, we used the similarity ratings and the embedding space to evaluate how well several popular models conform to human similarity judgments. One finding is that more complex models that perform better on task-specific benchmarks do not better conform to human semantic judgments. In addition to the human similarity judgments, pre-trained psychological embeddings and code for inferring variational embeddings are made publicly available. ImageNet-HSJ supports the appraisal of internal representations and the development of more human-like models.
****
What's in the Image? Explorable Decoding of Compressed Images
Yuval Bahat, Tomer Michaeli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2908-2917
The ever-growing amounts of visual contents captured on a daily basis necessitate the use of lossy compression methods in order to save storage space and transmission bandwidth. While extensive research efforts are devoted to improving compression techniques, every method inevitably discards information. Especially at low bit rates, this information often corresponds to semantically meaningful visual cues, so that decompression involves significant ambiguity. In spite of this fact, existing decompression algorithms typically produce only a single output, and do not allow the viewer to explore the set of images that map to the given compressed code. In this work we propose the first image decompression method to facilitate user-exploration of the diverse set of natural images that could have given rise to the compressed input code, thus granting users the ability to determine what could and what could not have been there in the original scene. Specifically, we develop a novel deep-network based decoder architecture for the ubiquitous JPEG standard, which allows traversing the set of decompressed images that are consistent with the compressed JPEG file. To allow for simple user interaction, we develop a graphical user interface comprising several intuitive exploration tools, including an automatic tool for examining specific solutions of interest. We exemplify our framework on graphical, medical and forensic use cases, demonstrating its wide range of potential applications.
****
Context Modeling in 3D Human Pose Estimation: A Unified Perspective
Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6238-6247
Estimating 3D human pose from a single image suffers from severe ambiguity since

multiple 3D joint configurations may have the same 2D projection. The state-of-the-art methods often rely on context modeling methods such as pictorial structure model (PSM) or graph neural network (GNN) to reduce ambiguity. However, there is no study that rigorously compares them side by side. So we first present a general formula for context modeling in which both PSM and GNN are its special cases. By comparing the two methods, we found that the end-to-end training scheme in GNN and the limb length constraints in PSM are two complementary factors to improve results. To combine their advantages, we propose ContextPose based on attention mechanism that allows enforcing soft limb length constraints in a deep network. The approach effectively reduces the chance of getting absurd 3D pose estimates with incorrect limb lengths and achieves state-of-the-art results on two benchmark datasets. More importantly, the introduction of limb length constraints into deep networks enables the approach to achieve much better generalization performance.
****

Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, Jingjing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7331-7341

The canonical approach to video-and-language learning (e.g., video question answering) dictates a neural model to learn from offline-extracted dense video features from vision models and text features from language models. These feature extractors are trained independently and usually on tasks different from the target domains, rendering these fixed features sub-optimal for downstream tasks. Moreover, due to the high computational overload of dense video features, it is often difficult (or infeasible) to plug feature extractors directly into existing approaches for easy finetuning. To provide a remedy to this dilemma, we propose a generic framework CLIPBERT that enables affordable end-to-end learning for video-and-language tasks, by employing sparse sampling, where only a single or a few sparsely sampled short clips from a video are used at each training step. Experiments on text-to-video retrieval and video question answering on six datasets demonstrate that CLIPBERT outperforms (or is on par with) existing methods that exploit full-length videos, suggesting that end-to-end learning with just a few sparsely sampled clips is often more accurate than using densely extracted offline features from full-length videos, proving the proverbial less-is-more principle. Videos in the datasets are from considerably different domains and lengths, ranging from 3-second generic-domain GIF videos to 180-second YouTube human activity videos, showing the generalization ability of our approach. Comprehensive ablation studies and thorough analyses are provided to dissect what factors lead to this success.
****

Consensus Maximisation Using Influences of Monotone Boolean Functions

Ruwan Tennakoon, David Suter, Erchuan Zhang, Tat-Jun Chin, Alireza Bab-Hadiashar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2866-2875

Consensus maximisation (MaxCon), widely used for robust fitting in computer vision, aims to find the largest subset of data that fits the model within some tolerance level. In this paper, we outline the connection between MaxCon problem and the abstract problem of finding the maximum upper zero of a Monotone Boolean Function (MBF) defined over the Boolean Cube. Then, we link the concept of influences (in a MBF) to the concept of outlier (in MaxCon) and show that influences of points belonging to the largest structure in data would be the smallest under certian conditions. Based on this observation, we present an iterative algorithm to perform consensus maximisation. Results for both synthetic and real visual data experiments show that the MBF based algorithm is capable of generating a near optimal solution relatively quickly. This is particularly important where there are large number of outliers (gross or pseudo) in the observed data.
****

Meta-Mining Discriminative Samples for Kinship Verification

Wanhua Li, Shiwei Wang, Jiwen Lu, Jianjiang Feng, Jie Zhou; Proceedings of the I

EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16135-16144
Kinship verification aims to find out whether there is a kin relation for a given pair of facial images. Kinship verification databases are born with unbalanced data. For a database with N positive kinship pairs, we naturally obtain N(N-1) negative pairs. How to fully utilize the limited positive pairs and mine discriminative information from sufficient negative samples for kinship verification remains an open issue. To address this problem, we propose a Discriminative Sample Meta-Mining (DSMM) approach in this paper. Unlike existing methods that usually construct a balanced dataset with fixed negative pairs, we propose to utilize all possible pairs and automatically learn discriminative information from data. Specifically, we sample an unbalanced train batch and a balanced meta-train batch for each iteration. Then we learn a meta-miner with the meta-gradient on the balanced meta-train batch. In the end, the samples in the unbalanced train batch are re-weighted by the learned meta-miner to optimize the kinship models. Experimental results on the widely used KinFaceW-I, KinFaceW-II, TSKinFace, and Cornell Kinship datasets demonstrate the effectiveness of the proposed approach.
****
AQD: Towards Accurate Quantized Object Detection
Peng Chen, Jing Liu, Bohan Zhuang, Mingkui Tan, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 104-113
Network quantization allows inference to be conducted using low-precision arithmetic for improved inference efficiency of deep neural networks on edge devices. However, designing aggressively low-bit (e.g., 2-bit) quantization schemes on complex tasks, such as object detection, still remains challenging in terms of severe performance degradation and unverifiable efficiency on common hardware. In this paper, we propose an Accurate Quantized object Detection solution, termed AQD, to fully get rid of floating-point computation. To this end, we target using fixed-point operations in all kinds of layers, including the convolutional layers, normalization layers, and skip connections, allowing the inference to be executed using integer-only arithmetic. To demonstrate the improved latency-vs-accuracy trade-off, we apply the proposed methods on RetinaNet and FCOS. In particular, experimental results on MS-COCO dataset show that our AQD achieves comparable or even better performance compared with the full-precision counterpart under extremely low-bit schemes, which is of great practical value. Source code and models are available at: https://github.com/aim-uofa/model-quantization
****
Learning Cross-Modal Retrieval With Noisy Labels
Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, Jie Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5403-5413
Recently, cross-modal retrieval is emerging with the help of deep multimodal learning. However, even for unimodal data, collecting large-scale well-annotated data is expensive and time-consuming, and not to mention the additional challenges from multiple modalities. Although crowd-sourcing annotation, e.g., Amazon's Mechanical Turk, can be utilized to mitigate the labeling cost, but leading to the unavoidable noise in labels for the non-expert annotating. To tackle the challenge, this paper presents a general Multimodal Robust Learning framework (MRL) for learning with multimodal noisy labels to mitigate noisy samples and correlate distinct modalities simultaneously. To be specific, we propose a Robust Clustering loss (RC) to make the deep networks focus on clean samples instead of noisy ones. Besides, a simple yet effective multimodal loss function, called Multimodal Contrastive loss (MC), is proposed to maximize the mutual information between different modalities, thus alleviating the interference of noisy samples and cross-modal discrepancy. Extensive experiments are conducted on four widely-used multimodal datasets to demonstrate the effectiveness of the proposed approach by comparing to 14 state-of-the-art methods.
****
LOHO: Latent Optimization of Hairstyles via Orthogonalization

Rohit Saha, Brendan Duke, Florian Shkurti, Graham W. Taylor, Parham Aarabi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1984-1993
Hairstyle transfer is challenging due to hair structure differences in the source and target hair. Therefore, we propose Latent Optimization of Hairstyles via Orthogonalization (LOHO), an optimization-based approach using GAN inversion to infill missing hair structure details in latent space during hairstyle transfer. Our approach decomposes hair into three attributes: perceptual structure, appearance, and style, and includes tailored losses to model each of these attributes independently. Furthermore, we propose two-stage optimization and gradient orthogonalization to enable disentangled latent space optimization of our hair attributes. Using LOHO for latent space manipulation, users can synthesize novel photorealistic images by manipulating hair attributes either individually or jointly, transferring the desired attributes from reference hairstyles. LOHO achieves a superior FID compared with the current state-of-the-art (SOTA) for hairstyle transfer. Additionally, LOHO preserves the subject's identity comparably well according to PSNR and SSIM when compared to SOTA image embedding pipelines.
****
Single-Shot Freestyle Dance Reenactment
Oran Gafni, Oron Ashual, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 882-891
The task of motion transfer between a source dancer and a target person is a special case of the pose transfer problem, in which the target person changes their pose in accordance with the motions of the dancer. In this work, we propose a novel method that can reanimate a single image by arbitrary video sequences, unseen during training. The method combines three networks: (i) a segmentation-mapping network, (ii) a realistic frame-rendering network, and (iii) a face refinement network. By separating this task into three stages, we are able to attain a novel sequence of realistic frames, capturing natural motion and appearance. Our method obtains significantly better visual quality than previous methods and is able to animate diverse body types and appearances, which are captured in challenging poses.
****
A Quasiconvex Formulation for Radial Cameras
Carl Olsson, Viktor Larsson, Fredrik Kahl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14576-14585
In this paper we study structure from motion problems for 1D radial cameras. Under this model the projection of a 3D point is a line in the image plane going through the principal point, which makes the model invariant to radial distortion and changes in focal length. It can therefore effectively be applied to uncalibrated image collections without the need for explicit estimation of camera intrinsics. We show that the reprojection errors of 1D radial cameras are examples of quasiconvex functions. This opens up the possibility to solve a general class of relevant reconstruction problems globally optimally using tools from convex optimization. In fact, our resulting algorithm is based on solving a series of LP problems. We perform an extensive experimental evaluation, on both synthetic and real data, showing that a whole class of multiview geometry problems across a range of different cameras models with varying and unknown intrinsic calibration can be reliably and accurately solved within the same framework.
****
Self-Supervised Learning of Depth Inference for Multi-View Stereo
Jiayu Yang, Jose M. Alvarez, Miaomiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7526-7534
Recent supervised multi-view depth estimation networks have achieved promising results. Similar to all supervised approaches, these networks require ground-truth data during training. However, collecting a large amount of multi-view depth data is very challenging. Here, we propose a self-supervised learning framework for multi-view stereo that exploit pseudo labels from the input data. We start by learning to estimate depth maps as initial pseudo labels under an unsupervised learning framework relying on image reconstruction loss as supervision. We then

refine the initial pseudo labels using a carefully designed pipeline leveraging depth information inferred from a higher resolution image and neighboring views. We use these high-quality pseudo labels as the supervision signal to train the network and improve, iteratively, its performance by self-training. Extensive experiments on the DTU dataset show that our proposed self-supervised learning framework outperforms existing unsupervised multi-view stereo networks by a large margin and performs on par compared to the supervised counterpart. Code is available at https://github.com/JiayuYANG/Self-supervised-CVP-MVSNet
****
BRepNet: A Topological Message Passing System for Solid Models
Joseph G. Lambourne, Karl D.D. Willis, Pradeep Kumar Jayaraman, Aditya Sanghi, Peter Meltzer, Hooman Shayani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12773-12782
Boundary representation (B-rep) models are the standard way 3D shapes are described in Computer-Aided Design (CAD) applications. They combine lightweight parametric curves and surfaces with topological information which connects the geometric entities to describe manifolds. In this paper we introduce BRepNet, a neural network architecture designed to operate directly on B-rep data structures, avoiding the need to approximate the model as meshes or point clouds. BRepNet defines convolutional kernels with respect to oriented coedges in the data structure. In the neighborhood of each coedge, a small collection of faces, edges and coedges can be identified and patterns in the feature vectors from these entities detected by specific learnable parameters. In addition, to encourage further deep learning research with B-reps, we publish the Fusion 360 Gallery segmentation dataset. A collection of over 35,000 B-rep models annotated with information about the modeling operations which created each face. We demonstrate that BRepNet can segment these models with higher accuracy than methods working on meshes, and point clouds.
****
Learning To Predict Visual Attributes in the Wild
Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13018-13028
Visual attributes constitute a large portion of information contained in a scene. Objects can be described using a wide variety of attributes which portray their visual appearance (color, texture), geometry (shape, size, posture), and other intrinsic properties (state, action). Existing work is mostly limited to study of attribute prediction in specific domains. In this paper, we introduce a large-scale in-the-wild visual attribute prediction dataset consisting of over 927K attribute annotations for over 260K object instances. Formally, object attribute prediction is a multi-label classification problem where all attributes that apply to an object must be predicted. Our dataset poses significant challenges to existing methods due to large number of attributes, label sparsity, data imbalance, and object occlusion. To this end, we propose several techniques that systematically tackle these challenges, including a base model that utilizes both low- and high-level CNN features with multi-hop attention, reweighting and resampling techniques, a novel negative label expansion scheme, and a novel supervised attribute-aware contrastive learning algorithm. Using these techniques, we achieve near 3.7 mAP and 5.7 overall F1 points improvement over the current state of the art. Further details about the VAW dataset can be found at https://vawdataset.com/.
****
Animating Pictures With Eulerian Motion Fields
Aleksander Holynski, Brian L. Curless, Steven M. Seitz, Richard Szeliski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5810-5819
In this paper, we demonstrate a fully automatic method for converting a still image into a realistic animated looping video. We target scenes with continuous fluid motion, such as flowing water and billowing smoke. Our method relies on the observation that this type of natural motion can be convincingly reproduced from

a static Eulerian motion description, i.e. a single, temporally constant flow field that defines the immediate motion of a particle at a given 2D location. We use an image-to-image translation network to encode motion priors of natural scenes collected from online videos, so that for a new photo, we can synthesize a corresponding motion field. The image is then animated using the generated motion through a deep warping technique: pixels are encoded as deep features, those features are warped via Eulerian motion, and the resulting warped feature maps are decoded as images. In order to produce continuous, seamlessly looping video textures, we propose a novel video looping technique that flows features both forward and backward in time and then blends the results. We demonstrate the effectiveness and robustness of our method by applying it to a large collection of examples including beaches, waterfalls, and flowing rivers.
****
Generalized Focal Loss V2: Learning Reliable Localization Quality Estimation for Dense Object Detection
Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11632-11641
Localization Quality Estimation (LQE) is crucial and popular in the recent advancement of dense object detectors since it can provide accurate ranking scores that benefit the Non-Maximum Suppression processing and improve detection performance. As a common practice, most existing methods predict LQE scores through vanilla convolutional features shared with object classification or bounding box regression. In this paper, we explore a completely novel and different perspective to perform LQE -- based on the learned distributions of the four parameters of the bounding box. The bounding box distributions are inspired and introduced as "General Distribution" in GFLV1, which describes the uncertainty of the predicted bounding boxes well. Such a property makes the distribution statistics of a bounding box highly correlated to its real localization quality. Specifically, a bounding box distribution with a sharp peak usually corresponds to high localization quality, and vice versa. By leveraging the close correlation between distribution statistics and the real localization quality, we develop a considerably lightweight Distribution-Guided Quality Predictor (DGQP) for reliable LQE based on GFLV1, thus producing GFLV2. To our best knowledge, it is the first attempt in object detection to use a highly relevant, statistical representation to facilitate LQE. Extensive experiments demonstrate the effectiveness of our method. Notably, GFLV2 (ResNet-101) achieves 46.2 AP at 14.6 FPS, surpassing the previous state-of-the-art ATSS baseline (43.6 AP at 14.6 FPS) by absolute 2.6 AP on COCO \tt test-dev , without sacrificing the efficiency both in training and inference.
****
Cross-Domain Adaptive Clustering for Semi-Supervised Domain Adaptation
Jichang Li, Guanbin Li, Yemin Shi, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2505-2514
In semi-supervised domain adaptation, a few labeled samples per class in the target domain guide features of the remaining target samples to aggregate around them. However, the trained model cannot produce a highly discriminative feature representation for the target domain because the training data is dominated by labeled samples from the source domain. This could lead to disconnection between the labeled and unlabeled target samples as well as misalignment between unlabeled target samples and the source domain. In this paper, we propose a novel approach called Cross-domain Adaptive Clustering to address this problem. To achieve both inter-domain and intra-domain adaptation, we first introduce an adversarial adaptive clustering loss to group features of unlabeled target data into clusters and perform cluster-wise feature alignment across the source and target domains. We further apply pseudo labeling to unlabeled samples in the target domain and retain pseudo-labels with high confidence. Pseudo labeling expands the number of "labeled" samples in each class in the target domain, and thus produces a more robust and powerful cluster core for each class to facilitate adversarial learning. Extensive experiments on benchmark datasets, including DomainNet, Office-Home and Office, demonstrate that our proposed approach achieves the state-of-the-

art performance in semi-supervised domain adaptation.
****
ST3D: Self-Training for Unsupervised Domain Adaptation on 3D Object Detection
Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10368-10378
We present a new domain adaptive self-training pipeline, named ST3D, for unsupervised domain adaptation on 3D object detection from point clouds. First, we pre-train the 3D detector on the source domain with our proposed random object scaling strategy for mitigating the negative effects of source domain bias. Then, the detector is iteratively improved on the target domain by alternatively conducting two steps, which are the pseudo label updating with the developed quality-aware triplet memory bank and the model training with curriculum data augmentation. These specific designs for 3D object detection enable the detector to be trained with consistent and high-quality pseudo labels and to avoid overfitting to the large number of easy examples in pseudo labeled data. Our ST3D achieves state-of-the-art performance on all evaluated datasets and even surpasses fully supervised results on KITTI 3D object detection benchmark. Code will be available at https://github.com/CVMI-Lab/ST3D.
****
HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching
Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, Sofien Bouaziz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14362-14372
This paper presents HITNet, a novel neural network architecture for real-time stereo matching. Contrary to many recent neural network approaches that operate on a full costvolume and rely on 3D convolutions, our approach does not explicitly build a volume and instead relies on a fast multi-resolution initialization step, differentiable 2D geometric propagation and warping mechanisms to infer disparity hypotheses. To achieve a high level of accuracy, our network not only geometrically reasons about disparities but also infers slanted plane hypotheses allowing to more accurately perform geometric warping and upsampling operations. Our architecture is inherently multi-resolution allowing the propagation of information across different levels. Multiple experiments prove the effectiveness of the proposed approach at a fraction of the computation required by the state-of-the-art methods. At the time of writing, HITNet ranks 1st-3rd on all the metrics published on the ETH3D website for two view stereo, ranks 1st on most of the metrics amongst all the end-to-end learning approaches on Middleburyv3, ranks 1st on the popular KITTI 2012 and 2015 benchmarks among the published methods faster than 100ms.
****
VaB-AL: Incorporating Class Imbalance and Difficulty With Variational Bayes for Active Learning
Jongwon Choi, Kwang Moo Yi, Jihoon Kim, Jinho Choo, Byoungjip Kim, Jinyeop Chang, Youngjune Gwon, Hyung Jin Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6749-6758
Active Learning for discriminative models has largely been studied with the focus on individual samples, with less emphasis on how classes are distributed or which classes are hard to deal with. In this work, we show that this is harmful. We propose a method based on the Bayes' rule, that can naturally incorporate class imbalance into the Active Learning framework. We derive that three terms should be considered together when estimating the probability of a classifier making a mistake for a given sample; i) probability of mislabelling a class, ii) likelihood of the data given a predicted class, and iii) the prior probability on the abundance of a predicted class. Implementing these terms requires a generative model and an intractable likelihood estimation. Therefore, we train a Variational Auto Encoder (VAE) for this purpose. To further tie the VAE with the classifier and facilitate VAE training, we use the classifiers' deep feature representations as input to the VAE. By considering all three probabilities, among them espec

ially the data imbalance, we can substantially improve the potential of existing methods under limited data budget. We show that our method can be applied to classification tasks on multiple different datasets -- including one that is a real-world dataset with heavy data imbalance -- significantly outperforming the state of the art.
****
Exploiting & Refining Depth Distributions With Triangulation Light Curtains
Yaadhav Raaj, Siddharth Ancha, Robert Tamburo, David Held, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7434-7442
Active sensing through the use of Adaptive Depth Sensors is a nascent field, with potential in areas such as Advanced driver-assistance systems (ADAS). They do however require dynamically driving a laser / light-source to a specific location to capture information, with one such class of sensor being the Triangulation Light Curtains (LC). In this work, we introduce a novel approach that exploits prior depth distributions from RGB cameras to drive a Light Curtain's laser line to regions of uncertainty to get new measurements. These measurements are utilized such that depth uncertainty is reduced and errors get corrected recursively. We show real-world experiments that validate our approach in outdoor and driving settings, and demonstrate qualitative and quantitative improvements in depth RMSE when RGB cameras are used in tandem with a Light Curtain.
****
DG-Font: Deformable Generative Networks for Unsupervised Font Generation
Yangchen Xie, Xinyuan Chen, Li Sun, Yue Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5130-5140
Font generation is a challenging problem especially for some writing systems that consist of a large number of characters and has attracted a lot of attention in recent years. However, existing methods for font generation are often in supervised learning. They require a large number of paired data, which is labor-intensive and expensive to collect. Besides, common image-to-image translation models often define style as the set of textures and colors, which cannot be directly applied to font generation. To address these problems, we propose novel deformable generative networks for unsupervised font generation (DG-Font). We introduce a feature deformation skip connection (FDSC) which predicts pairs of displacement maps and employs the predicted maps to apply deformable convolution to the low-level feature maps from the content encoder. The outputs of FDSC are fed into a mixer to generate the final results. Taking advantage of FDSC, the mixer outputs a high-quality character with a complete structure. To further improve the quality of generated images, we use three deformable convolution layers in the content encoder to learn style-invariant feature representations. Experiments demonstrate that our model generates characters in higher quality than state-of-art methods. The source code is available at https://github.com/ecnuycxie/DG-Font.
****
Deep Multi-Task Learning for Joint Localization, Perception, and Prediction
John Phillips, Julieta Martinez, Ioan Andrei Barsan, Sergio Casas, Abbas Sadat, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4679-4689
Over the last few years, we have witnessed tremendous progress on many subtasks of autonomous driving including perception, motion forecasting, and motion planning. However, these systems often assume that the car is accurately localized against a high-definition map. In this paper we question this assumption, and investigate the issues that arise in state-of-the-art autonomy stacks under localization error. Based on our observations, we design a system that jointly performs perception, prediction, and localization. Our architecture is able to reuse computation between the three tasks, and is thus able to correct localization errors efficiently. We show experiments on a large-scale autonomy dataset, demonstrating the efficiency and accuracy of our proposed approach.
****
Deeply Shape-Guided Cascade for Instance Segmentation
Hao Ding, Siyuan Qiao, Alan Yuille, Wei Shen; Proceedings of the IEEE/CVF Confer

ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8278-8288
The key to a successful cascade architecture for precise instance segmentation is to fully leverage the relationship between bounding box detection and mask segmentation across multiple stages. Although modern instance segmentation cascades achieve leading performance, they mainly make use of a unidirectional relationship, i.e., mask segmentation can benefit from iteratively refined bounding box detection. In this paper, we investigate an alternative direction, i.e., how to take the advantage of precise mask segmentation for bounding box detection in a cascade architecture. We propose a Deeply Shape-guided Cascade (DSC) for instance segmentation, which iteratively imposes the shape guidances extracted from mask prediction at previous stage on bounding box detection at current stage. It forms a bi-directional relationship between the two tasks by introducing three key components: (1) Initial shape guidance: A mask-supervised Region Proposal Network (mPRN) with the ability to generate class-agnostic masks; (2) Explicit shape guidance: A mask-guided region-of-interest (RoI) feature extractor, which employs mask segmentation at previous stage to focus feature extraction at current stage within a region aligned well with the shape of the instance-of-interest rather than a rectangular RoI; (3) Implicit shape guidance: A feature fusion operation which feeds intermediate mask features at previous stage to the bounding box head at current stage. Experimental results show that DSC outperforms the state-of-the-art instance segmentation cascade, Hybrid Task Cascade (HTC), by a large margin and achieves 51.8 box AP and 45.5 mask AP on COCO test-dev. The code is released at: https://github.com/hding2455/DSC.
****
MetricOpt: Learning To Optimize Black-Box Evaluation Metrics
Chen Huang, Shuangfei Zhai, Pengsheng Guo, Josh Susskind; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 174-183
We study the problem of directly optimizing arbitrary non-differentiable task evaluation metrics such as misclassification rate and recall. Our method, named MetricOpt, operates in a black-box setting where the computational details of the target metric are unknown. We achieve this by learning a differentiable value function, which maps compact task-specific model parameters to metric observations. The learned value function is easily pluggable into existing optimizers like SGD and Adam, and is effective for rapidly finetuning a pre-trained model. This leads to consistent improvements since the value function provides effective metric supervision during finetuning, and helps to correct the potential bias of loss-only supervision. MetricOpt achieves state-of-the-art performance on a variety of metrics for (image) classification, image retrieval and object detection. Solid benefits are found over competing methods, which often involve complex loss design or adaptation. MetricOpt also generalizes well to new tasks and model architectures.
****
Multispectral Photometric Stereo for Spatially-Varying Spectral Reflectances: A Well Posed Problem?
Heng Guo, Fumio Okura, Boxin Shi, Takuya Funatomi, Yasuhiro Mukaigawa, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 963-971
Multispectral photometric stereo (MPS) aims at recovering the surface normal of a scene from a single-shot multispectral image, which is known as an ill-posed problem. To make the problem well-posed, existing MPS methods rely on restrictive assumptions, such as shape prior, surfaces having a monochromatic with uniform albedo. This paper alleviates the restrictive assumptions in existing methods. We show that the problem becomes well-posed for a surface with a uniform chromaticity but spatially-varying albedos based on our new formulation. Specifically, if at least three (or two) scene points share the same chromaticity, the proposed method uniquely recovers their surface normals and spectral reflectance with the illumination of more than or equal to four (or five) spectral lights. Besides, our method can be made robust by having many (i.e., 4 or more) spectral bands using robust estimation techniques for conventional photometric stereo. Experimen

ts on both synthetic and real-world scenes demonstrate the effectiveness of our method. Our data and result can be found at https://github.com/GH-HOME/MultispectralPS.git.
****
Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback
Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, Rogerio Feris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11307-11317
Conversational interfaces for the detail-oriented retail fashion domain are more natural, expressive, and user friendly than classical keyword-based search interfaces. In this paper, we introduce the Fashion IQ dataset to support and advance research on interactive fashion image retrieval. Fashion IQ is the first fashion dataset to provide human-generated captions that distinguish similar pairs of garment images together with side-information consisting of real-world product descriptions and derived visual attribute labels for these images. We provide a detailed analysis of the characteristics of the Fashion IQ data, and present a transformer-based user simulator and interactive image retriever that can seamlessly integrate visual attributes with image features, user feedback, and dialog history, leading to improved performance over the state of the art in dialog-based image retrieval. We believe that our dataset will encourage further work on developing more natural and real-world applicable conversational shopping assistants.
****
Few-Shot Human Motion Transfer by Personalized Geometry and Texture Modeling
Zhichao Huang, Xintong Han, Jia Xu, Tong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2297-2306
We present a new method for few-shot human motion transfer that achieves realistic human image generation with only a small number of appearance inputs. Despite recent advances in single person motion transfer, prior methods often require a large number of training images and take long training time. One promising direction is to perform few-shot human motion transfer, which only needs a few of source images for appearance transfer. However, it is particularly challenging to obtain satisfactory transfer results. In this paper, we address this issue by rendering a human texture map to a surface geometry (represented as a UV map), which is personalized to the source person. Our geometry generator combines the shape information from source images, and the pose information from 2D keypoints to synthesize the personalized UV map. A texture generator then generates the texture map conditioned on the texture of source images to fill out invisible parts. Furthermore, we may fine-tune the texture map on the manifold of the texture generator from a few source images at the test time, which improves the quality of the texture map without over-fitting or artifacts. Extensive experiments show the proposed method outperforms state-of-the-art methods both qualitatively and quantitatively. Our code is available at https://github.com/HuangZhiChao95/FewShotMotionTransfer.
****
HDMapGen: A Hierarchical Graph Generative Model of High Definition Maps
Lu Mi, Hang Zhao, Charlie Nash, Xiaohan Jin, Jiyang Gao, Chen Sun, Cordelia Schmid, Nir Shavit, Yuning Chai, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4227-4236
High Definition (HD) maps are maps with precise definitions of road lanes with rich semantics of the traffic rules. They are critical for several key stages in an autonomous driving system, including motion forecasting and planning. However, there are only a small amount of real-world road topologies and geometries, which significantly limits our ability to test out the self-driving stack to generalize onto new unseen scenarios. To address this issue, we introduce a new challenging task to generate HD maps. In this work, we explore several autoregressive models using different data representations, including sequence, plain graph, and hierarchical graph. We propose HDMapGen, a hierarchical graph generation model capable of producing high-quality and diverse HD maps through a coarse-to-fine approach. Experiments on the Argoverse dataset and an in-house dataset show tha

t HDMapGen significantly outperforms baseline methods. Additionally, we demonstrate that HDMapGen achieves high efficiency and scalability.
****
GeoSim: Realistic Video Simulation via Geometry-Aware Composition for Self-Driving
Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchen Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7230-7240
Scalable sensor simulation is an important yet challenging open problem for safety-critical domains such as self-driving. Current works in image simulation either fail to be photorealistic or do not model the 3D environment and the dynamic objects within, losing high-level control and physical realism. In this paper, we present GeoSim, a geometry-aware image composition process which synthesizes novel urban driving scenarios by augmenting existing images with dynamic objects extracted from other scenes and rendered at novel poses. Towards this goal, we first build a diverse bank of 3D objects with both realistic geometry and appearance from sensor data. During simulation, we perform a novel geometry-aware simulation-by-composition procedure which 1) proposes plausible and realistic object placements into a given scene, 2) render novel views of dynamic objects from the asset bank, and 3) composes and blends the rendered image segments. The resulting synthetic images are realistic, traffic-aware, and geometrically consistent, allowing our approach to scale to complex use cases. We demonstrate two such important applications: long-range realistic video simulation across multiple camera sensors, and synthetic data generation for data augmentation on downstream segmentation tasks. Please check https://tmux.top/publication/geosim/ for high-resolution video results.
****
AlphaMatch: Improving Consistency for Semi-Supervised Learning With Alpha-Divergence
Chengyue Gong, Dilin Wang, Qiang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13683-13692
Semi-supervised learning (SSL) is a key approach toward more data-efficient machine learning by jointly leverage both labeled and unlabeled data. We propose AlphaMatch, an efficient SSL method that leverages data augmentations, by efficiently enforcing the label consistency between the data points and the augmented data derived from them. Our key technical contribution lies on: 1) using alpha-divergence to prioritize the regularization on data with high Semi-supervised learning (SSL) is a key approach toward more data-efficient machine learning by jointly leverage both labeled and unlabeled data. We propose AlphaMatch, an efficient SSL method that leverages data augmentations, by efficiently enforcing the label consistency between the data points and the augmented data derived from them. Our key technical contribution lies on: 1) using alpha-divergence to prioritize the regularization on data with high confidence, achieving similar effect as FixMatch but in a more flexible fashion, and 2) proposing an optimization-based, EM-like algorithm to enforce the consistency, which enjoys better convergence than iterative regularization procedures used in recent SSL methods such as FixMatch, UDA, and MixMatch. AlphaMatch is simple and easy to implement, and consistently outperforms prior arts on standard benchmarks, e.g. CIFAR-10, SVHN, CIFAR-100, STL-10. Specifically, we achieve 91.3% test accuracy on CIFAR-10 with just 4 labelled data per class, substantially improving over the previously best 88.7% accuracy achieved by FixMatch.
****
Unbalanced Feature Transport for Exemplar-Based Image Translation
Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, Chunyan Miao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15028-15038
Despite the great success of GANs in images translation with different conditioned inputs such as semantic segmentation and edge map, generating high-fidelity i

mages with reference styles from exemplars remains a grand challenge in conditio nal image-to-image translation. This paper presents a general image translation framework that incorporates optimal transport for feature alignment between cond itional inputs and style exemplars in translation. The introduction of optimal t ransport mitigates the constraint of many-to-one feature matching significantly while building up semantic correspondences between conditional inputs and exempl ars. We design a novel unbalanced optimal transport to address the transport bet ween features with deviational distributions which exists widely between conditi onal inputs and exemplars. In addition, we design a semantic-aware normalization scheme that injects style and semantic features of exemplars into the image tra nslation process successfully. Extensive experiments over multiple image transla tion tasks show that our proposed technique achieves superior image translation qualitatively and quantitatively as compared with the state-of-the-art.
****

Self-Generated Defocus Blur Detection via Dual Adversarial Discriminators
Wenda Zhao, Cai Shang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2021, pp. 6933-6942
Although existing fully-supervised defocus blur detection (DBD) models significa ntly improve performance, training such deep models requires abundant pixel-leve l manual annotation, which is highly time-consuming and error-prone. Addressing this issue, this paper makes an effort to train a deep DBD model without using a ny pixel-level annotation. The core insight is that a defocus blur region/focuse d clear area can be arbitrarily pasted to a given realistic full blurred image/f ull clear image without affecting the judgment of the full blurred image/full cl ear image. Specifically, we train a generator G in an adversarial manner against dual discriminators Dc and Db. G learns to produce a DBD mask that generates a composite clear image and a composite blurred image through copying the focused area and unfocused region from corresponding source image to another full clear image and full blurred image. Then, Dc and Db can not distinguish them from real istic full clear image and full blurred image simultaneously, achieving a self-g enerated DBD by an implicit manner to define what a defocus blur area is. Beside s, we propose a bilateral triplet-excavating constraint to avoid the degenerate problem caused by the case one discriminator defeats the other one. Comprehensiv e experiments on two widely-used DBD datasets demonstrate the superiority of the proposed approach. Source codes are available at: https://github.com/shangcai1/ SG.
****

View Generalization for Single Image Textured 3D Models
Anand Bhattad, Aysegul Dundar, Guilin Liu, Andrew Tao, Bryan Catanzaro; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2021, pp. 6081-6090
Humans can easily infer the underlying 3D geometry and texture of an object only from a single 2D image. Current computer vision methods can do this, too, but s uffer from view generalization problems -- the models inferred tend to make poor predictions of appearance in novel views. As for generalization problems in mac hine learning, the difficulty is balancing single-view accuracy (cf. training er ror; bias) with novel view accuracy (cf. test error; variance). We describe a cl ass of models whose geometric rigidity is easily controlled to manage this trade off. We describe a cycle consistency loss that improves view generalization (rou ghly, a model from a generated view should predict the original view well). View generalization of textures requires that models share texture information, so a car seen from the back still has headlights because other cars have headlights. We describe a cycle consistency loss that encourages model textures to be align ed, so as to encourage sharing. We compare our method against the state-of-the-a rt method and show both qualitative and quantitative improvements.
****

Your "Flamingo" is My "Bird": Fine-Grained, or Not
Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, Jun Guo; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11476-11485

Whether what you see in Figure 1 is a "flamingo" or a "bird", is the question we ask in this paper. While fine-grained visual classification (FGVC) strives to arrive at the former, for the majority of us non-experts just "bird" would probably suffice. The real question is therefore -- how can we tailor for different fine-grained definitions under divergent levels of expertise. For that, we re-envisage the traditional setting of FGVC, from single-label classification, to that of top-down traversal of a pre-defined coarse-to-fine label hierarchy -- so that our answer becomes "bird"="Phoenicopteriformes"="Phoenicopteridae"="flamingo". To approach this new problem, we first conduct a comprehensive human study where we confirm that most participants prefer multi-granularity labels, regardless whether they consider themselves experts. We then discover the key intuition that: coarse-level label prediction exacerbates fine-grained feature learning, yet fine-level feature betters the learning of coarse-level classifier. This discovery enables us to design a very simple albeit surprisingly effective solution to our new problem, where we (i) leverage level-specific classification heads to disentangle coarse-level features with fine-grained ones, and (ii) allow finer-grained features to participate in coarser-grained label predictions, which in turn helps with better disentanglement. Experiments show that our method achieves superior performance in the new FGVC setting, and performs better than state-of-the-art on traditional single-label FGVC problem as well. Thanks to its simplicity, our method can be easily implemented on top of any existing FGVC frameworks and is parameter-free. Codes are available at: https://github.com/PRIS-CV/Fine-Grained-or-Not
****
Anchor-Constrained Viterbi for Set-Supervised Action Segmentation
Jun Li, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9806-9815
This paper is about action segmentation under weak supervision in training, where the ground truth provides only a set of actions present, but neither their temporal ordering nor when they occur in a training video. We use a Hidden Markov Model (HMM) grounded on a multilayer perceptron (MLP) to label video frames, and thus generate a pseudo-ground truth for the subsequent pseudo-supervised training. In testing, a Monte Carlo sampling of action sets seen in training is used to generate candidate temporal sequences of actions, and select the maximum posterior sequence. Our key contribution is a new anchor-constrained Viterbi algorithm (ACV) for generating the pseudo-ground truth, where anchors are salient action parts estimated for each action from a given ground-truth set. Our evaluation on the tasks of action segmentation and alignment on the benchmark Breakfast, MPII Cooking2, Hollywood Extended datasets demonstrates our superior performance relative to that of prior work.
****
SOON: Scenario Oriented Object Navigation With Graph-Based Exploration
Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12689-12699
The ability to navigate like a human towards a language-guided target from anywhere in a 3D embodied environment is one of the 'holy grail' goals of intelligent robots. Most visual navigation benchmarks, however, focus on navigating toward a target from a fixed starting point, guided by an elaborate set of instructions that depicts step-by-step. This approach deviates from real-world problems in which human-only describes what the object and its surrounding look like and asks the robot to start navigation from anywhere. Accordingly, in this paper, we introduce a Scenario Oriented Object Navigation (SOON) task. In this task, an agent is required to navigate from an arbitrary position in a 3D embodied environment to localize a target following a scene description. To give a promising direction to solve this task, we propose a novel graph-based exploration (GBE) method, which models the navigation state as a graph and introduces a novel graph-based exploration approach to learn knowledge from the graph and stabilize training by learning sub-optimal trajectories. We also propose a new large-scale benchmark named From Anywhere to Object (FAO) dataset. To avoid target ambiguity, the desc

riptions in FAO provide rich semantic scene information includes: object attribute, object relationship, region description, and nearby region description. Our experiments reveal that the proposed GBE outperforms various state-of-the-arts on both FAO and R2R datasets. And the ablation studies on FAO validates the quality of the dataset.
****
Learning Scalable lY=-Constrained Near-Lossless Image Compression via Joint Lossy Image and Residual Compression
Yuanchao Bai, Xianming Liu, Wangmeng Zuo, Yaowei Wang, Xiangyang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11946-11955
We propose a novel joint lossy image and residual compression framework for learning l_infinity-constrained near-lossless image compression. Specifically, we obtain a lossy reconstruction of the raw image through lossy image compression and uniformly quantize the corresponding residual to satisfy a given tight l_infinity error bound. Suppose that the error bound is zero, i.e., lossless image compression, we formulate the joint optimization problem of compressing both the lossy image and the original residual in terms of variational auto-encoders and solve it with end-to-end training. To achieve scalable compression with the error bound larger than zero, we derive the probability model of the quantized residual by quantizing the learned probability model of the original residual, instead of training multiple networks. We further correct the bias of the derived probability model caused by the context mismatch between training and inference. Finally, the quantized residual is encoded according to the bias-corrected probability model and is concatenated with the bitstream of the compressed lossy image. Experimental results demonstrate that our near-lossless codec achieves the state-of-the-art performance for lossless and near-lossless image compression, and achieves competitive PSNR while much smaller l_infinity error compared with lossy image codecs at high bit rates.
****
Minimally Invasive Surgery for Sparse Neural Networks in Contrastive Manner
Chong Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3589-3598
With the development of deep learning, neural networks tend to be deeper and larger to achieve good performance. Trained models are more compute-intensive and memory-intensive, which lead to the big challenges on memory bandwidth, storage, latency, and throughput. In this paper, we propose the neural network compression method named minimally invasive surgery. Different from traditional model compression and knowledge distillation methods, the proposed method refers to the minimally invasive surgery principle. It learns the principal features from a pair of dense and compressed models in a contrastive manner. It also optimizes the neural networks to meet the specific hardware acceleration requirements. Through qualitative, quantitative, and ablation experiments, the proposed method shows a compelling performance, acceleration, and generalization in various tasks.
****
XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations
Eunji Kim, Siwon Kim, Minji Seo, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15719-15728
Automated diagnosis using deep neural networks in chest radiography can help radiologists detect life-threatening diseases. However, existing methods only provide predictions without accurate explanations, undermining the trustworthiness of the diagnostic methods. Here, we present XProtoNet, a globally and locally interpretable diagnosis framework for chest radiography. XProtoNet learns representative patterns of each disease from X-ray images, which are prototypes, and makes a diagnosis on a given X-ray image based on the patterns. It predicts the area where a sign of the disease is likely to appear and compares the features in the predicted area with the prototypes. It can provide a global explanation, the prototype, and a local explanation, how the prototype contributes to the prediction of a single image. Despite the constraint for interpretability, XProtoNet achieves state-of-the-art classification performance on the public NIH chest X-ray d

ataset.
****

Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution

Baoli Sun, Xinchen Ye, Baopu Li, Haojie Li, Zhihui Wang, Rui Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7792-7801

Existing color-guided depth super-resolution (DSR) approaches require paired RGB-D data as training examples where the RGB image is used as structural guidance to recover the degraded depth map due to their geometrical similarity. However, the paired data may be limited or expensive to be collected in actual testing environment. Therefore, we explore for the first time to learn the cross-modal knowledge at training stage, where both RGB and depth modalities are available, but test on the target dataset, where only single depth modality exists. Our key idea is to distill the knowledge of scene structural guidance from color modality to the single DSR task without changing its network architecture. Specifically, we propose an auxiliary depth estimation (DE) task that takes color image as input to estimate a depth map, and train both DSR task and DE task collaboratively to boost the performance of DSR. A cross-task distillation module is designed to realize bilateral cross-task knowledge transfer. Moreover, to address the problem of RGB-D structure inconsistency and boost the structure perception, we advance a structure prediction (SP) task that provides extra structure regularization to help both DSR and DE networks learn more informative structure representations for depth recovery. Extensive experiments demonstrate that our scheme achieves superior performance in comparison with other DSR methods.
****

Visual Navigation With Spatial Attention

Bar Mayo, Tamir Hazan, Ayellet Tal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16898-16907

This work focuses on object goal visual navigation, aiming at finding the location of an object from a given class, where in each step the agent is provided with an egocentric RGB image of the scene. We propose to learn the agent's policy using a reinforcement learning algorithm. Our key contribution is a novel attention probability model for visual navigation tasks. This attention encodes semantic information about observed objects, as well as spatial information about their place. This combination of the "what"" and the "where"" allows the agent to navigate toward the sought-after object effectively. The attention model is shown to improve the agent's policy and to achieve state-of-the-art results on commonly-used datasets.
****

Model-Based 3D Hand Reconstruction via Self-Supervised Learning

Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10451-10460

Reconstructing a 3D hand from a single-view RGB image is challenging due to various hand configurations and depth ambiguity. To reliably reconstruct a 3D hand from a monocular image, most state-of-the-art methods heavily rely on 3D annotations at the training stage, but obtaining 3D annotations is expensive. To alleviate reliance on labeled training data, we propose S2HAND, a self-supervised 3D hand reconstruction network that can jointly estimate pose, shape, texture, and the camera viewpoint. Specifically, we obtain geometric cues from the input image through easily accessible 2D detected keypoints. To learn an accurate hand reconstruction model from these noisy geometric cues, we utilize the consistency between 2D and 3D representations and propose a set of novel losses to rationalize outputs of the neural network. For the first time, we demonstrate the feasibility of training an accurate 3D hand reconstruction network without relying on manual annotations. Our experiments show that the proposed method achieves comparable performance with recent fully-supervised methods while using fewer supervision data.
****

Robust Reflection Removal With Reflection-Free Flash-Only Cues
Chenyang Lei, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14811-14820
We propose a simple yet effective reflection-free cue for robust reflection removal from a pair of flash and ambient (no-flash) images. The reflection-free cue exploits a flash-only image obtained by subtracting the ambient image from the corresponding flash image in raw data space. The flash-only image is equivalent to an image taken in a dark environment with only a flash on. We observe that this flash-only image is visually reflection-free, and thus it can provide robust cues to infer the reflection in the ambient image. Since the flash-only image usually has artifacts, we further propose a dedicated model that not only utilizes the reflection-free cue but also avoids introducing artifacts, which helps accurately estimate reflection and transmission. Our experiments on real-world images with various types of reflection demonstrate the effectiveness of our model with reflection-free flash-only cues: our model outperforms state-of-the-art reflection removal approaches by more than 5.23dB in PSNR, 0.04 in SSIM, and 0.068 in LPIPS. Our source code and dataset are publicly available at github.com/ChenyangLEI/flash-reflection-removal.
****
Real-Time Selfie Video Stabilization
Jiyang Yu, Ravi Ramamoorthi, Keli Cheng, Michel Sarkis, Ning Bi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12036-12044
We propose a novel real-time selfie video stabilization method. Our method is completely automatic and runs at 26 fps. We use a 1D linear convolutional network to directly infer the rigid moving least squares warping which implicitly balances between the global rigidity and local flexibility. Our network structure is specifically designed to stabilize the background and foreground at the same time, while providing optional control of stabilization focus (relative importance of foreground vs. background) to the users. To train our network, we collect a selfie video dataset with 1005 videos, which is significantly larger than previous selfie video datasets. We also propose a grid approximation to the rigid moving least squares that enables the real-time frame warping. Our method is fully automatic and produces visually and quantitatively better results than previous real-time general video stabilization methods. Compared to previous offline selfie video methods, our approach produces comparable quality with a speed improvement of orders of magnitude. Our code and selfie video dataset is available at https://github.com/jiy173/selfievideostabilization.
****
3D Human Action Representation Learning via Cross-View Consistency Pursuit
Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4741-4750
In this work, we propose a Cross-view Contrastive Learning framework for unsupervised 3D skeleton-based action representation (CrosSCLR), by leveraging multi-view complementary supervision signal. CrosSCLR consists of both single-view contrastive learning (SkeletonCLR) and cross-view consistent knowledge mining (CVC-KM) modules, integrated in a collaborative learning manner. It is noted that CVC-KM works in such a way that high-confidence positive/negative samples and their distributions are exchanged among views according to their embedding similarity, ensuring cross-view consistency in terms of contrastive context, i.e., similar distributions. Extensive experiments show that CrosSCLR achieves remarkable action recognition results on NTU-60 and NTU-120 datasets under unsupervised settings, with observed higher-quality action representations. Our code is available at https://github.com/LinguoLi/CrosSCLR.
****
Differentiable SLAM-Net: Learning Particle SLAM for Visual Navigation
Peter Karkus, Shaojun Cai, David Hsu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2815-2825
Simultaneous localization and mapping (SLAM) remains challenging for a number of

downstream applications, such as visual robot navigation, because of rapid turns, featureless walls, and poor camera quality. We introduce the Differentiable SLAM Network (SLAM-net) along with a navigation architecture to enable planar robot navigation in previously unseen indoor environments. SLAM-net encodes a particle filter based SLAM algorithm in a differentiable computation graph, and learns task-oriented neural network components by backpropagating through the SLAM algorithm. Because it can optimize all model components jointly for the end-objective, SLAM-net learns to be robust in challenging conditions. We run experiments in the Habitat platform with different real-world RGB and RGB-D datasets. SLAM-net significantly outperforms the widely adapted ORB-SLAM in noisy conditions. Our navigation architecture with SLAM-net improves the state-of-the-art for the Habitat Challenge 2020 PointNav task by a large margin (37% to 64% success).
****
Learning Goals From Failure
Dave Epstein, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11194-11204
We introduce a framework that predicts the goals behind observable human action in video. Motivated by evidence in developmental psychology, we leverage video of unintentional action to learn video representations of goals without direct supervision. Our approach models videos as contextual trajectories that represent both low-level motion and high-level action features. Experiments and visualizations show our trained model is able to predict the underlying goals in video of unintentional action. We also propose a method to "automatically correct" unintentional action by leveraging gradient signals of our model to adjust latent trajectories. Although the model is trained with minimal supervision, it is competitive with or outperforms baselines trained on large (supervised) datasets of successfully executed goals, showing that observing unintentional action is crucial to learning about goals in video.
****
Rank-One Prior: Toward Real-Time Scene Recovery
Jun Liu, Wen Liu, Jianing Sun, Tieyong Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14802-14810
Scene recovery is a fundamental imaging task for several practical applications, e.g., video surveillance and autonomous vehicles, etc. To improve visual quality under different weather/imaging conditions, we propose a real-time light correction method to recover the degraded scenes in the cases of sandstorms, underwater, and haze. The heart of our work is that we propose an intensity projection strategy to estimate the transmission. This strategy is motivated by a straightforward rank-one transmission prior. The complexity of transmission estimation is O(N) where N is the size of the single image. Then we can recover the scene in real-time. Comprehensive experiments on different types of weather/imaging conditions illustrate that our method outperforms competitively several state-of-the-art imaging methods in terms of efficiency and robustness.
****
Body2Hands: Learning To Infer 3D Hands From Conversational Gesture Body Dynamics
Evonne Ng, Shiry Ginosar, Trevor Darrell, Hanbyul Joo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11865-11874
We propose a novel learned deep prior of body motion for 3D hand shape synthesis and estimation in the domain of conversational gestures. Our model builds upon the insight that body motion and hand gestures are strongly correlated in non-verbal communication settings. We formulate the learning of this prior as a prediction task of 3D hand shape over time given body motion input alone. Trained with 3D pose estimations obtained from a large-scale dataset of internet videos, our hand prediction model produces convincing 3D hand gestures given only the 3D motion of the speaker's arms as input. We demonstrate the efficacy of our method on hand gesture synthesis from body motion input, and as a strong body prior for single-view image-based 3D hand pose estimation. We demonstrate that our method outperforms previous state-of-the-art approaches and can generalize beyond the monologue-based training data to multi-person conversations.

****
Linear Semantics in Generative Adversarial Networks
Jianjin Xu, Changxi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9351-9360
Generative Adversarial Networks (GANs) are able to generate high-quality images, but it remains difficult to explicitly specify the semantics of synthesized images. In this work, we aim to better understand the semantic representation of GANs, and thereby enable semantic control in GAN's generation process. Interestingly, we find that a well-trained GAN encodes image semantics in its internal feature maps in a surprisingly simple way: a linear transformation of feature maps suffices to extract the generated image semantics. To verify this simplicity, we conduct extensive experiments on various GANs and datasets; and thanks to this simplicity, we are able to learn a semantic segmentation model for a trained GAN from a small number (e.g., 8) of labeled images. Last but not least, leveraging our finding, we propose two few-shot image editing approaches, namely Semantic-Conditional Sampling and Semantic Image Editing. Given a trained GAN and as few as eight semantic annotations, the user is able to generate diverse images subject to a user-provided semantic layout, and control the synthesized image semantics. We have made the code publicly available.
****
Mesoscopic Photogrammetry With an Unstabilized Phone Camera
Kevin C. Zhou, Colin Cooke, Jaehee Park, Ruobing Qian, Roarke Horstmeyer, Joseph A. Izatt, Sina Farsiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7535-7545
We present a feature-free photogrammetric technique that enables quantitative 3D mesoscopic (mm-scale height variation) imaging with tens-of-micron accuracy from sequences of images acquired by a smartphone at close range (several cm) under freehand motion without additional hardware. Our end-to-end, pixel-intensity-based approach jointly registers and stitches all the images by estimating a coaligned height map, which acts as a pixel-wise radial deformation field that orthorectifies each camera image to allow plane-plus-parallax registration. The height maps themselves are reparameterized as the output of an untrained encoder-decoder convolutional neural network (CNN) with the raw camera images as the input, which effectively removes many reconstruction artifacts. Our method also jointly estimates both the camera's dynamic 6D pose and its distortion using a nonparametric model, the latter of which is especially important in mesoscopic applications when using cameras not designed for imaging at short working distances, such as smartphone cameras. We also propose strategies for reducing computation time and memory, applicable to other multi-frame registration problems. Finally, we demonstrate our method using sequences of multi-megapixel images captured by an unstabilized smartphone on a variety of samples (e.g., painting brushstrokes, circuit board, seeds).
****
Joint Generative and Contrastive Learning for Unsupervised Person Re-Identification
Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, Francois Bremond; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2004-2013
Recent self-supervised contrastive learning provides an effective approach for unsupervised person re-identification (ReID) by learning invariance from different views (transformed versions) of an input. In this paper, we incorporate a Generative Adversarial Network (GAN) and a contrastive learning module into one joint training framework. While the GAN provides online data augmentation for contrastive learning, the contrastive module learns view-invariant features for generation. In this context, we propose a mesh-based view generator. Specifically, mesh projections serve as references towards generating novel views of a person. In addition, we propose a view-invariant loss to facilitate contrastive learning between original and generated views. Deviating from previous GAN-based unsupervised ReID methods involving domain adaptation, we do not rely on a labeled source dataset, which makes our method more flexible. Extensive experimental results s

how that our method significantly outperforms state-of-the-art methods under bot h, fully unsupervised and unsupervised domain adaptive settings on several large scale ReID datsets.
****

Wide-Baseline Multi-Camera Calibration Using Person Re-Identification
Yan Xu, Yu-Jhe Li, Xinshuo Weng, Kris Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13134-13143
We address the problem of estimating the 3D pose of a network of cameras for large-environment wide-baseline scenarios, e.g., cameras for construction sites, sports stadiums, and public spaces. This task is challenging since detecting and matching the same 3D keypoint observed from two very different camera views is difficult, making standard structure-from-motion (SfM) pipelines inapplicable. In such circumstances, treating people in the scene as "keypoints" and associating them across different camera views can be an alternative method for obtaining correspondences. Based on this intuition, we propose a method that uses ideas from person re-identification (re-ID) for wide-baseline camera calibration. Our method first employs a re-ID method to associate human bounding boxes across cameras, then converts bounding box correspondences to point correspondences, and finally solves for camera pose using multi-view geometry and bundle adjustment. Since our method does not require specialized calibration targets except for visible people, it applies to situations where frequent calibration updates are required. We perform extensive experiments on datasets captured from scenes of different sizes, camera settings (indoor and outdoor), and human activities (walking, playing basketball, construction). Experiment results show that our method achieves similar performance to standard SfM methods relying on manually labeled point correspondences.
****

ATSO: Asynchronous Teacher-Student Optimization for Semi-Supervised Image Segmentation
Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1235-1244
Semi-supervised learning is a useful tool for image segmentation, mainly due to its ability in extracting knowledge from unlabeled data to assist learning from labeled data. This paper focuses on a popular pipeline known as self-learning, where we point out a weakness named lazy mimicking that refers to the inertia that a model retains the prediction from itself and thus resists updates. To alleviate this issue, we propose the Asynchronous Teacher-Student Optimization (ATSO) algorithm that (i) breaks up continual learning from teacher to student and (ii) partitions the unlabeled training data into two subsets and alternately uses one subset to fine-tune the model which updates the labels on the other. We show the ability of ATSO on medical and natural image segmentation. In both scenarios, our method reports competitive performance, on par with the state-of-the-arts, in either using partial labeled data in the same dataset or transferring the trained model to an unlabeled dataset.
****

Panoramic Image Reflection Removal
Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C. Kot, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7762-7771
This paper studies the problem of panoramic image reflection removal, aiming at reliving the content ambiguity between reflection and transmission scenes. Although a partial view of the reflection scene is included in the panoramic image, it cannot be utilized directly due to its misalignment with the reflection-contaminated image. We propose a two-step approach to solve this problem, by first accomplishing geometric and photometric alignment for the reflection scene via a coarse-to-fine strategy, and then restoring the transmission scene via a recovery network. The proposed method is trained with a synthetic dataset and verified quantitatively with a real panoramic image dataset. The effectiveness of the proposed method is validated by the significant performance advantage over single ima

ge-based reflection removal methods and generalization capacity to limited-FoV s
cenarios captured by conventional camera or mobile phone users.
****
OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations
Yang Tan, Yang Li, Shao-Lun Huang; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2021, pp. 15779-15788
Transfer learning across heterogeneous data distributions (a.k.a. domains) and d
istinct tasks is a more general and challenging problem than conventional transf
er learning, where either domains or tasks are assumed to be the same. While neu
ral network based feature transfer is widely used in transfer learning applicati
ons, finding the optimal transfer strategy still requires time-consuming experim
ents and domain knowledge. We propose a transferability metric called Optimal Tr
ansport based Conditional Entropy (OTCE), to analytically predict the transfer p
erformance for supervised classification tasks in such cross-domain and cross-ta
sk feature transfer settings. Our OTCE score characterizes transferability as a
combination of domain difference and task difference, and explicitly evaluates t
hem from data in a unified framework. Specifically, we use optimal transport to
estimate domain difference and the optimal coupling between source and target di
stributions, which is then used to derive the conditional entropy of the target
task (task difference). Experiments on the largest cross-domain dataset DomainNe
t and Office31 demonstrate that OTCE shows an average of 21% gain in the correla
tion with the ground truth transfer accuracy compared to state-of-the-art method
s. We also investigate two applications of the OTCE score including source model
 selection and multi-source feature fusion.
****
Diverse Semantic Image Synthesis via Probability Distribution Modeling
Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua,
Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2021, pp. 7962-7971
Semantic image synthesis, translating semantic layouts to photo-realistic images
, is a one-to-many mapping problem. Though impressive progress has been recently
 made, diverse semantic synthesis that can efficiently produce semantic-level mu
ltimodal results, still remains a challenge. In this paper, we propose a novel d
iverse semantic image synthesis framework from the perspective of semantic class
 distributions, which naturally supports diverse generation at semantic or even
instance level. We achieve this by modeling class-level conditional modulation p
arameters as continuous probability distributions instead of discrete values, an
d sampling per-instance modulation parameters through instance-adaptive stochast
ic sampling that is consistent across the network. Moreover, we propose prior no
ise remapping, through linear perturbation parameters encoded from paired refere
nces, to facilitate supervised training and exemplar-based instance style contro
l at test time. Extensive experiments on multiple datasets show that our method
can achieve superior diversity and comparable quality compared to state-of-the-a
rt methods. Code will be available at https://github.com/tzt101/INADE.git
****
NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections
Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Al
exey Dosovitskiy, Daniel Duckworth; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2021, pp. 7210-7219
We present a learning-based method for synthesizingnovel views of complex scenes
 using only unstructured collections of in-the-wild photographs. We build on Neu
ral Radiance Fields (NeRF), which uses the weights of a multi-layer perceptron t
o model the density and color of a scene as a function of 3D coordinates. While
NeRF works well on images of static subjects captured under controlled settings,
 it is incapable of modeling many ubiquitous, real-world phenomena in uncontroll
ed images, such as variable illumination or transient occluders. We introduce a
series of extensions to NeRF to address these issues, thereby enabling accurate
reconstructions from unstructured image collections taken from the internet. We
apply our system, dubbed NeRF-W, to internet photo collections of famous landmar
ks,and demonstrate temporally consistent novel view renderings that are signific

antly closer to photorealism than the prior state of the art.
****
Learning by Watching
Jimuyang Zhang, Eshed Ohn-Bar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12711-12721
When in a new situation or geographical location, human drivers have an extraordinary ability to watch others and learn maneuvers that they themselves may have never performed. In contrast, existing techniques for learning to drive preclude such a possibility as they assume direct access to an instrumented ego-vehicle with fully known observations and expert driver actions. However, such measurements cannot be directly accessed for the non-ego vehicles when learning by watching others. Therefore, in an application where data is regarded as a highly valuable asset, current approaches completely discard the vast portion of the training data that can be potentially obtained through indirect observation of surrounding vehicles. Motivated by this key insight, we propose the Learning by Watching (LbW) framework which enables learning a driving policy without requiring full knowledge of neither the state nor expert actions. To increase its data, i.e., with new perspectives and maneuvers, LbW makes use of the demonstrations of other vehicles in a given scene by (1) transforming the ego-vehicle's observations to their points of view, and (2) inferring their expert actions. Our LbW agent learns more robust driving policies while enabling data-efficient learning, including quick adaptation of the policy to rare and novel scenarios. In particular, LbW drives robustly even with a fraction of available driving data required by existing methods, achieving an average success rate of 92% on the original CARLA benchmark with only 30 minutes of total driving data and 82% with only 10 minutes.
****
Pseudo Facial Generation With Extreme Poses for Face Recognition
Guoli Wang, Jiaqi Ma, Qian Zhang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1994-2003
Face recognition has achieved a great success in recent years, it is still challenging to recognize those facial images with extreme poses. Traditional methods consider it as a domain gap problem. Many of them settle it by generating fake frontal faces from extreme ones, whereas they are tough to maintain the identity information with high computational consumption and uncontrolled disturbances. Our experimental analysis shows a dramatic precision drop with extreme poses. Meanwhile, those extreme poses just exist minor visual differences after small rotations. Derived from this insight, we attempt to relieve such a huge precision drop by making minor changes to the input images without modifying existing discriminators. A novel lightweight pseudo facial generation is proposed to relieve the problem of extreme poses without generating any frontal facial image. It can depict the facial contour information and make appropriate modifications to preserve the critical identity information. Specifically, the proposed method reconstructs pseudo profile faces by minimizing the pixel-wise differences with original profile faces and maintaining the identity consistent information from their corresponding frontal faces simultaneously. The proposed framework can improve existing discriminators and obtain a great promotion on several benchmark datasets.
****
Inverting Generative Adversarial Renderer for Face Reconstruction
Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15619-15628
Given a monocular face image as input, 3D face geometry reconstruction aims to recover a corresponding 3Dface mesh. Recently, both optimization-based and learning-based face reconstruction methods have taken advantage of the emerging differentiable renderer and shown promising results. However, the differentiable renderer, mainly based on graphics rules, simplifies the realistic mechanism of the illumination, reflection, etc., of the real world, thus can-not produce realistic images. This brings a lot of domain-shift noise to the optimization or training

process. In this work, we introduce a novel Generative Adversarial Renderer (GAR) and propose to tailor its inverted version to the general fitting pipeline, to tackle the above problem. Specifically, the carefully designed neural renderer takes a face normal map and a latent code representing other fac-tors as inputs and renders a realistic face image. Since the GAR learns to model the complicated real-world image, instead of relying on the simplified graphics rules, it is capable of producing realistic images, which essentially inhibits the domain-shift noise in training and optimization. Equipped with the elaborated GAR, we further proposed a novel approach to predict 3D face parameters, in which we first obtain fine initial parameters via Renderer Invertingand then refine it with gradient-based optimizers. Extensive experiments have been conducted to demonstrate the effectiveness of the proposed generative adversarial renderer and the novel optimization-based face reconstruction framework. Our method achieves state-of-the-art performance on multiple face reconstruction datasets.
****
Efficient Object Embedding for Spliced Image Retrieval
Bor-Chun Chen, Zuxuan Wu, Larry S. Davis, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14965-14975
Detecting spliced images is one of the emerging challenges in computer vision. Unlike prior methods that focus on detecting low-level artifacts generated during the manipulation process, we use an image retrieval approach to tackle this problem. When given a spliced query image, our goal is to retrieve the original image from a database of authentic images. To achieve this goal, we propose representing an image by its constituent objects based on the intuition that the finest granularity of manipulations is oftentimes at the object-level. We introduce a framework, object embeddings for spliced image retrieval (OE-SIR), that utilizes modern object detectors to localize object regions. Each region is then embedded and collectively used to represent the image. Further, we propose a student-teacher training paradigm for learning discriminative embeddings within object regions to avoid expensive multiple forward passes. Detailed analysis of the efficacy of different feature embedding models is also provided in this study. Extensive experimental results show that the OE-SIR achieves state-of-the-art performance in spliced image retrieval.
****
GrooMeD-NMS: Grouped Mathematically Differentiable NMS for Monocular 3D Object Detection
Abhinav Kumar, Garrick Brazil, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8973-8983
Modern 3D object detectors have immensely benefited from the end-to-end learning idea. However, most of them use a post-processing algorithm called Non-Maximal Suppression (NMS) only during inference. While there were attempts to include NMS in the training pipeline for tasks such as 2D object detection, they have been less widely adopted due to a non-mathematical expression of the NMS. In this paper, we present and integrate GrooMeD-NMS -- a novel Grouped Mathematically Differentiable NMS for monocular 3D object detection, such that the network is trained end-to-end with a loss on the boxes after NMS. We first formulate NMS as a matrix operation and then group and mask the boxes in an unsupervised manner to obtain a simple closed-form expression of the NMS. GrooMeD-NMS addresses the mismatch between training and inference pipelines and, therefore, forces the network to select the best 3D box in a differentiable manner. As a result, GrooMeD-NMS achieves state-of-the-art monocular 3D object detection results on the KITTI benchmark dataset performing comparably to monocular video-based methods.
****
Flow Guided Transformable Bottleneck Networks for Motion Retargeting
Jian Ren, Menglei Chai, Oliver J. Woodford, Kyle Olszewski, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10795-10805
Human motion retargeting aims to transfer the motion of one person in a driving video or set of images to another person. Existing efforts leverage a long train

ing video from each target person to train a subject-specific motion transfer mo
del. However, the scalability of such methods is limited, as each model can only
 generate videos for the given target subject, and such training videos are labo
r-intensive to acquire and process. Few-shot motion transfer techniques, which o
nly require one or a few images from a target, have recently drawn considerable
attention. Methods addressing this task generally use either 2D or explicit 3D r
epresentations to transfer motion, and in doing so, sacrifice either accurate ge
ometric modeling or the flexibility of an end-to-end learned representation. Ins
pired by the Transformable Bottleneck Network, which renders novel views and man
ipulations of rigid objects, we propose an approach based on an implicit volumet
ric representation of the image content, which can then be spatially manipulated
 using volumetric flow fields. We address the challenging question of how to agg
regate information across different body poses, learning flow fields that allow
for combining content from the appropriate regions of input images of highly non
-rigid human subjects performing complex motions into a single implicit volumetr
ic representation. This allows us to learn our 3D representation solely from vid
eos of moving people. Armed with both 3D object understanding and end-to-end lea
rned rendering, this categorically novel representation delivers state-of-the-ar
t image generation quality, as shown by our quantitative and qualitative evaluat
ions.
****
Projecting Your View Attentively: Monocular Road Scene Layout Estimation via Cro
ss-View Transformation
Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, Jia Pan;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2021, pp. 15536-15545
HD map reconstruction is crucial for autonomous driving. LiDAR-based methods are
 limited due to the deployed expensive sensors and time-consuming computation. C
amera-based methods usually need to separately perform road segmentation and vie
w transformation, which often causes distortion and the absence of content. To p
ush the limits of the technology, we present a novel framework that enables reco
nstructing a local map formed by road layout and vehicle occupancy in the bird's
-eye view given a front-view monocular image only. In particular, we propose a c
ross-view transformation module, which takes the constraint of cycle consistency
 between views into account and makes full use of their correlation to strengthe
n the view transformation and scene understanding. Considering the relationship
between vehicles and roads, we also design a context-aware discriminator to furt
her refine the results. Experiments on public benchmarks show that our method ac
hieves the state-of-the-art performance in the tasks of road layout estimation a
nd vehicle occupancy estimation. Especially for the latter task, our model outpe
rforms all competitors by a large margin. Furthermore, our model runs at 35 FPS
on a single GPU, which is efficient and applicable for real-time panorama HD map
 reconstruction.
****
Deep Analysis of CNN-Based Spatio-Temporal Representations for Action Recognitio
n
Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John C
ohn, Aude Oliva, Quanfu Fan; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2021, pp. 6165-6175
In recent years, a number of approaches based on 2D or 3D convolutional neural n
etworks (CNN) have emerged for video action recognition, achieving state-of-the-
art results on several large-scale benchmark datasets. In this paper, we carry o
ut in-depth comparative analysis to better understand the differences between th
ese approaches and the progress made by them. To this end, we develop an unified
 framework for both 2D-CNN and 3D-CNN action models, which enables us to remove
bells and whistles and provides a common ground for fair comparison. We then con
duct an effort towards a large-scale analysis involving over 300 action recognit
ion models. Our comprehensive analysis reveals that a) a significant leap is mad
e in efficiency for action recognition, but not in accuracy; b) 2D-CNN and 3D-CN
N models behave similarly in terms of spatio-temporal representation abilities a

nd transferability. Our codes are available at https://github.com/IBM/action-recognition-pytorch.
****
Generalizable Person Re-Identification With Relevance-Aware Mixture of Experts
Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, Ling-Yu Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16145-16154
Domain generalizable (DG) person re-identification (ReID) is a challenging problem because we cannot access any unseen target domain data during training. Almost all the existing DG ReID methods follow the same pipeline where they use a hybrid dataset from multiple source domains for training, and then directly apply the trained model to the unseen target domains for testing. These methods often neglect individual source domains' discriminative characteristics and their relevances w.r.t. the unseen target domains, though both of which can be leveraged to help the model's generalization. To handle the above two issues, we propose a novel method called the relevance-aware mixture of experts (RaMoE), using an effective voting-based mixture mechanism to dynamically leverage source domains' diverse characteristics to improve the model's generalization. Specifically, we propose a decorrelation loss to make the source domain networks (experts) keep the diversity and discriminability of individual domains' characteristics. Besides, we design a voting network to adaptively integrate all the experts' features into the more generalizable aggregated features with domain relevance. Considering the target domains' invisibility during training, we propose a novel learning-to-learn algorithm combined with our relation alignment loss to update the voting network. Extensive experiments demonstrate that our proposed RaMoE outperforms the state-of-the-art methods.
****
Part-Aware Panoptic Segmentation
Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, Gijs Dubbelman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5485-5494
In this work, we introduce the new scene understanding task of Part-aware Panoptic Segmentation (PPS), which aims to understand a scene at multiple levels of abstraction, and unifies the tasks of scene parsing and part parsing. For this novel task, we provide consistent annotations on two commonly used datasets: Cityscapes and Pascal VOC. Moreover, we present a single metric to evaluate PPS, called Part-aware Panoptic Quality (PartPQ). For this new task, using the metric and annotations, we set multiple baselines by merging results of existing state-of-the-art methods for panoptic segmentation and part segmentation. Finally, we conduct several experiments that evaluate the importance of the different levels of abstraction in this single task.
****
Unsupervised Degradation Representation Learning for Blind Super-Resolution
Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10581-10590
Most existing CNN-based super-resolution (SR) methods are developed based on an assumption that the degradation is fixed and known (e.g., bicubic downsampling). However, these methods suffer a severe performance drop when the real degradation is different from their assumption. To handle various unknown degradations in real-world applications, previous methods rely on degradation estimation to reconstruct the SR image. Nevertheless, degradation estimation methods are usually time-consuming and may lead to SR failure due to large estimation errors. In this paper, we propose an unsupervised degradation representation learning scheme for blind SR without explicit degradation estimation. Specifically, we learn abstract representations to distinguish various degradations in the representation space rather than explicit estimation in the pixel space. Moreover, we introduce a Degradation-Aware SR (DASR) network with flexible adaption to various degradations based on the learned representations. It is demonstrated that our degradation representation learning scheme can extract discriminative representations to

obtain accurate degradation information. Experiments on both synthetic and real images show that our network achieves state-of-the-art performance for the blind SR task. Code is available at: https://github.com/LongguangWang/DASR.
****
Convolutional Hough Matching Networks
Juhong Min, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2940-2950
Despite advances in feature representation, leveraging geometric relations is crucial for establishing reliable visual correspondences under large variations of images. In this work we introduce a Hough transform perspective on convolutional matching and propose an effective geometric matching algorithm, dubbed Convolutional Hough Matching (CHM). The method distributes similarities of candidate matches over a geometric transformation space and evaluate them in a convolutional manner. We cast it into a trainable neural layer with a semi-isotropic high-dimensional kernel, which learns non-rigid matching with a small number of interpretable parameters. To validate the effect, we develop the neural network with CHM layers that perform convolutional matching in the space of translation and scaling. Our method sets a new state of the art on standard benchmarks for semantic visual correspondence, proving its strong robustness to challenging intra-class variations.
****
Hierarchical and Partially Observable Goal-Driven Policy Learning With Goals Relational Graph
Xin Ye, Yezhou Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14101-14110
We present a novel two-layer hierarchical reinforcement learning approach equipped with a Goals Relational Graph (GRG) for tackling the partially observable goal-driven task, such as goal-driven visual navigation. Our GRG captures the underlying relations of all goals in the goal space through a Dirichlet-categorical process that facilitates: 1) the high-level network raising a sub-goal towards achieving a designated final goal; 2) the low-level network towards an optimal policy; and 3) the overall system generalizing unseen environments and goals. We evaluate our approach with two settings of partially observable goal-driven tasks -- a grid-world domain and a robotic object search task. Our experimental results show that our approach exhibits superior generalization performance on both unseen environments and new goals.
****
Point 4D Transformer Networks for Spatio-Temporal Modeling in Point Cloud Videos
Hehe Fan, Yi Yang, Mohan Kankanhalli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14204-14213
Point cloud videos exhibit irregularities and lack of order along the spatial dimension where points emerge inconsistently across different frames. To capture the dynamics in point cloud videos, point tracking is usually employed. However, as points may flow in and out across frames, computing accurate point trajectories is extremely difficult. Moreover, tracking usually relies on point colors and thus may fail to handle colorless point clouds. In this paper, to avoid point tracking, we propose a novel Point 4D Transformer (P4Transformer) network to model raw point cloud videos. Specifically, P4Transformer consists of (i) a point 4D convolution to embed the spatio-temporal local structures presented in a point cloud video and (ii) a transformer to capture the appearance and motion information across the entire video by performing self-attention on the embedded local features. In this fashion, related or similar local areas are merged with attention weight rather than by explicit tracking. Extensive experiments, including 3D action recognition and 4D semantic segmentation, on four benchmarks demonstrate the effectiveness of our P4Transformer for point cloud video modeling.
****
CoCoNets: Continuous Contrastive 3D Scene Representations
Shamit Lal, Mihir Prabhudesai, Ishita Mediratta, Adam W. Harley, Katerina Fragkiadaki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12487-12496

This paper explores self-supervised learning of amodal 3D feature representations from RGB and RGB-D posed images and videos, agnostic to object and scene semantic content, and evaluates the resulting scene representations in the downstream tasks of visual correspondence, object tracking, and object detection. The model infers a latent 3D representation of the scene in the form of 3D feature points, where each continuous world 3D point is mapped to its corresponding feature vector. The model is trained for contrastive view prediction by rendering 3D feature clouds in queried viewpoints and matching against the 3D feature point cloud predicted from the query view. Notably, the representation can be queried for any 3D location, even if it is not visible from the input view. Our model brings together three powerful ideas of recent exciting research work: 3D feature grids as a neural bottleneck for view prediction, implicit functions for handling resolution limitations of 3D grids, and contrastive learning for unsupervised training of feature representations. We show the resulting 3D visual feature representations effectively scale across objects and scenes, imagine information occluded or missing from the input viewpoints, track objects over time, align semantically related objects in 3D, and improve 3D object detection. We outperform many existing state-of-the-art methods for 3D feature learning and view prediction, which are either limited by 3D grid spatial resolution, do not attempt to build amodal 3D representations, or do not handle combinatorial scene variability due to their non-convolutional bottlenecks.
****
Distribution Alignment: A Unified Framework for Long-Tail Visual Recognition
Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2361-2370
Despite the success of the deep neural networks, it remains challenging to effectively build a system for long-tail visual recognition tasks. To address this problem, we first investigate the performance bottleneck of the two-stage learning framework via ablative study. Motivated by our discovery, we develop a unified distribution alignment strategy for long-tail visual recognition. Particularly, we first propose an adaptive calibration strategy for each data point to calibrate its classification scores. Then we introduce a generalized re-weight method to incorporate the class prior, which provides a flexible and unified solution to copy with diverse scenarios of various visual recognition tasks. We validate our method by extensive experiments on four tasks, including image classification, semantic segmentation, object detection, and instance segmentation. Our approach achieves the state-of-the-art results across all four recognition tasks with a simple and unified framework.
****
Dynamic Class Queue for Large Scale Face Recognition in the Wild
Bi Li, Teng Xi, Gang Zhang, Haocheng Feng, Junyu Han, Jingtuo Liu, Errui Ding, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3763-3772
Learning discriminative representation using large-scale face datasets in the wild is crucial for real-world applications, yet it remains challenging. The difficulties lie in many aspects and this work focus on computing resource constraint and long-tailed class distribution. Recently, classification-based representation learning with deep neural networks and well-designed losses have demonstrated good recognition performance. However, the computing and memory cost linearly scales up to the number of identities (classes) in the training set, and the learning process suffers from unbalanced classes. In this work, we propose a dynamic class queue (DCQ) to tackle these two problems. Specifically, for each iteration during training, a subset of classes for recognition are dynamically selected and their class weights are dynamically generated on-the-fly which are stored in a queue. Since only a subset of classes is selected for each iteration, the computing requirement is reduced. By using a single server without model parallel, we empirically verify in large-scale datasets that 10% of classes are sufficient to achieve similar performance as using all classes. Moreover, the class weights are dynamically generated in a few-shot manner and therefore suitable for tail

classes with only a few instances. We show clear improvement over a strong baseline in the largest public dataset Megaface Challenge2 (MF2) which has 672K identities and over 88% of them have less than 10 instances. Code is available at https://github.com/bilylee/DCQ
****

3D-MAN: 3D Multi-Frame Attention Network for Object Detection
Zetong Yang, Yin Zhou, Zhifeng Chen, Jiquan Ngiam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1863-1872
3D object detection is an important module in autonomous driving and robotics. However, many existing methods focus on using single frames to perform 3D detection, and do not fully utilize information from multiple frames. In this paper, we present 3D-MAN: a 3D multi-frame attention network that effectively aggregates features from multiple perspectives and achieves state-of-the-art performance on Waymo Open Dataset. 3D-MAN first uses a novel fast single-frame detector to produce box proposals. The box proposals and their corresponding feature maps are then stored in a memory bank. We design a multi-view alignment and aggregation module, using attention networks, to extract and aggregate the temporal features stored in the memory bank. This effectively combines the features coming from different perspectives of the scene. We demonstrate the effectiveness of our approach on the large-scale complex Waymo Open Dataset, achieving state-of-the-art results compared to published single-frame and multi-frame methods.
****

Cross-Modal Center Loss for 3D Cross-Modal Retrieval
Longlong Jing, Elahe Vahdani, Jiaxing Tan, Yingli Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3142-3151
Cross-modal retrieval aims to learn discriminative and modal-invariant features for data from different modalities. Unlike the existing methods which usually learn from the features extracted by offline networks, in this paper, we propose an approach to jointly train the components of cross-modal retrieval framework with metadata, and enable the network to find optimal features. The proposed end-to-end framework is updated with three loss functions: 1) a novel cross-modal center loss to eliminate cross-modal discrepancy, 2) cross-entropy loss to maximize inter-class variations, and 3) mean-square-error loss to reduce modality variations. In particular, our proposed cross-modal center loss minimizes the distances of features from objects belonging to the same class across all modalities. Extensive experiments have been conducted on the retrieval tasks across multi-modalities including 2D image, 3D point cloud and mesh data. The proposed framework significantly outperforms the state-of-the-art methods for both cross-modal and in-domain retrieval for 3D objects on the ModelNet10 and ModelNet40 datasets.
****

Learning View Selection for 3D Scenes
Yifan Sun, Qixing Huang, Dun-Yu Hsiao, Li Guan, Gang Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14464-14473
Efficient 3D space sampling to represent an underlying3D object/scene is essential for 3D vision, robotics, and be-yond. A standard approach is to explicitly sample a densecollection of views and formulate it as a view selection prob-lem, or, more generally, a set cover problem. In this paper,we introduce a novel approach that avoids dense view sam-pling. The key idea is to learn a view prediction networkand a trainable aggregation module that takes the predictedviews as input and outputs an approximation of their genericscores (e.g., surface coverage, viewing angle from surfacenormals). This methodology allows us to turn the set coverproblem (or multi-view representation optimization) into acontinuous optimization problem. We then explain how toeffectively solve the induced optimization problem using con-tinuation, i.e., aggregating a hierarchy of smoothed scoringmodules. Experimental results show that our approach ar-rives at similar or better solutions with about 10 x speed upin running time, comparing with the standard methods.
****

FESTA: Flow Estimation via Spatial-Temporal Attention for Scene Point Clouds
Haiyan Wang, Jiahao Pang, Muhammad A. Lodhi, Yingli Tian, Dong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14173-14182
Scene flow depicts the dynamics of a 3D scene, which is critical for various applications such as autonomous driving, robot navigation, AR/VR, etc. Conventionally, scene flow is estimated from dense/regular RGB video frames. With the development of depth-sensing technologies, precise 3D measurements are available via point clouds which have sparked new research in 3D scene flow. Nevertheless, it remains challenging to extract scene flow from point clouds due to the sparsity and irregularity in typical point cloud sampling patterns. One major issue related to irregular sampling is identified as the randomness during point set abstraction/feature extraction---an elementary process in many flow estimation scenarios. A novel Spatial Abstraction with Attention (SA^2) layer is accordingly proposed to alleviate the unstable abstraction problem. Moreover, a Temporal Abstraction with Attention (TA^2) layer is proposed to rectify attention in temporal domain, leading to benefits with motions scaled in a larger range. Extensive analysis and experiments verified the motivation and significant performance gains of our method, dubbed as Flow Estimation via Spatial-Temporal Attention (FESTA), when compared to several state-of-the-art benchmarks of scene flow estimation.
****
Semi-Supervised Action Recognition With Temporal Contrastive Learning
Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, Abir Das; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10389-10399
Learning to recognize actions from only a handful of labeled videos is a challenging problem due to the scarcity of tediously collected activity labels. We approach this problem by learning a two-pathway temporal contrastive model using unlabeled videos at two different speeds leveraging the fact that changing video speed does not change an action. Specifically, we propose to maximize the similarity between encoded representations of the same video at two different speeds as well as minimize the similarity between different videos played at different speeds. This way we use the rich supervisory information in terms of `time' that is present in otherwise unsupervised pool of videos. With this simple yet effective strategy of manipulating video playback rates, we considerably outperform video extensions of sophisticated state-of-the-art semi-supervised image recognition methods across multiple diverse benchmark datasets and network architectures. Interestingly, our proposed approach benefits from out-of-domain unlabeled videos showing generalization and robustness. We also perform rigorous ablations and analysis to validate our approach. Project page: https://cvir.github.io/TCL/.
****
SG-Net: Spatial Granularity Network for One-Stage Video Instance Segmentation
Dongfang Liu, Yiming Cui, Wenbo Tan, Yingjie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9816-9825
Video instance segmentation (VIS) is a new and critical task in computer vision. To date, top-performing VIS methods extend the two-stage Mask R-CNN by adding a tracking branch, leaving plenty of room for improvement. In contrast, we approach the VIS task from a new perspective and propose a one-stage spatial granularity network (SG-Net). SG-Net demonstrates four advantages: 1) Our task heads (detection, segmentation, and tracking) are crafted interdependently so they can effectively share features and enjoy the joint optimization; 2) Each of our task predictions avoids using proposal-based RoI features, resulting in much reduced runtime complexity per instance; 3) Our mask prediction is dynamically performed on the sub-regions of each detected instance, leading to high-quality masks of fine granularity; 4) Our tracking head models objects' centerness movements for tracking, which effectively enhances the tracking robustness to different object appearances. In evaluation, we present state-of-the-art comparisons on the YouTube-VIS dataset. Extensive experiments demonstrate that our compact one-stage method can achieve improved performance in both accuracy and inference speed. We hope our SG-Net could serve as a simple yet strong baseline for the VIS task. Code

will be available.
****
Learned Initializations for Optimizing Coordinate-Based Neural Representations
Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, Ren Ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2846-2855
Coordinate-based neural representations have shown significant promise as an alternative to discrete, array-based representations for complex low dimensional signals. However, optimizing a coordinate-based network from randomly initialized weights for each new signal is inefficient. We propose applying standard meta-learning algorithms to learn the initial weight parameters for these fully-connected networks based on the underlying class of signals being represented (e.g., images of faces or 3D models of chairs). Despite requiring only a minor change in implementation, using these learned initial weights enables faster convergence during optimization and can serve as a strong prior over the signal class being modeled, resulting in better generalization when only partial observations of a given signal are available. We explore these benefits across a variety of tasks, including representing 2D images, reconstructing CT scans, and recovering 3D shapes and scenes from 2D image observations.
****
Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization
Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 464-474
Localizing persons and recognizing their actions from videos is a challenging task towards high-level video under-standing. Recent advances have been achieved by modeling direct pairwise relations between entities. In this paper, we take one step further, not only model direct relations between pairs but also take into account indirect higher-order relations established upon multiple elements. We propose to explicitly model the Actor-Context-Actor Relation, which is the relation between two actors based on their interactions with the context. To this end, we design an Actor-Context-Actor Relation Network (ACAR-Net) which builds upon a novel High-order Relation Reasoning Operator and an Actor-Context Feature Bank to enable indirect relation reasoning for spatio-temporal action localization. Experiments on AVA and UCF101-24 datasets show the advantages of modeling actor-context-actor relations, and visualization of attention maps further verifies that our model is capable of finding relevant higher-order relations to support action detection. Notably, our method ranks first in the AVA-Kinetics action localization task of ActivityNet Challenge 2020, outperforming other entries by a significant margin (+6.71 mAP). The code is available online.
****
Cross-View Cross-Scene Multi-View Crowd Counting
Qi Zhang, Wei Lin, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 557-567
Multi-view crowd counting has been previously proposed to utilize multi-cameras to extend the field-of-view of a single camera, capturing more people in the scene, and improve counting performance for occluded people or those in low resolution. However, the current multi-view paradigm trains and tests on the same single scene and camera-views, which limits its practical application. In this paper, we propose a cross-view cross-scene (CVCS) multi-view crowd counting paradigm, where the training and testing occur on different scenes with arbitrary camera layouts. To dynamically handle the challenge of optimal view fusion under scene and camera layout change and non-correspondence noise due to camera calibration errors or erroneous features, we propose a CVCS model that attentively selects and fuses multiple views together using camera layout geometry, and a noise view regularization method to train the model to handle non-correspondence errors. We also generate a large synthetic multi-camera crowd counting dataset with a large number of scenes and camera views to capture many possible variations, which avoids the difficulty of collecting and annotating such a large real dataset. We then test our trained CVCS model on real multi-view counting datasets, by using u

nsupervised domain transfer. The proposed CVCS model trained on synthetic data o utperforms the same model trained only on real data, and achieves promising perf ormance compared to fully supervised methods that train and test on the same sin gle scene.
****
Semantic Segmentation With Generative Models: Semi-Supervised Learning and Stron g Out-of-Domain Generalization
Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, Sanja Fidler; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2021, pp. 8300-8311
Training deep networks with limited labeled data while achieving a strong genera lization ability is key in the quest to reduce human annotation efforts. This is the goal of semi-supervised learning, which exploits more widely available unla beled data to complement small labeled data sets. In this paper, we propose a no vel framework for discriminative pixel-level tasks using a generative model of b oth images and labels. Concretely, we learn a generative adversarial network tha t captures the joint image-label distribution and is trained efficiently using a large set of unlabeled images supplemented with only few labeled ones. We build our architecture on top of StyleGAN2, augmented with a label synthesis branch. Image labeling at test time is achieved by first embedding the target image into the joint latent space via an encoder network and test-time optimization, and t hen generating the label from the inferred embedding. We evaluate our approach i n two important domains: medical image segmentation and part-based face segmenta tion. We demonstrate strong in-domain performance compared to several baselines, and are the first to showcase extreme out-of-domain generalization, such as tra nsferring from CT to MRI in medical imaging, and photographs of real faces to pa intings, sculptures, and even cartoons and animal faces.
****
Depth-Aware Mirror Segmentation
Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, Xiaopeng We i; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2021, pp. 3044-3053
We present a novel mirror segmentation method that leverages depth estimates fro m ToF-based cameras as an additional cue to disambiguate challenging cases where the contrast or relation in RGB colors between the mirror reflection and the su rrounding scene is subtle. A key observation is that ToF depth estimates do not report the true depth of the mirror surface, but instead return the total length of the reflected light paths, thereby creating obvious depth discontinuities at the mirror boundaries. To exploit depth information in mirror segmentation, we first construct a large-scale RGB-D mirror segmentation dataset, which we subseq uently employ to train a novel depth-aware mirror segmentation framework. Our mi rror segmentation framework first locates the mirrors based on color and depth d iscontinuities and correlations. Next, our model further refines the mirror boun daries through contextual contrast taking into account both color and depth info rmation. We extensively validate our depth-aware mirror segmentation method and demonstrate that our model outperforms state-of-the-art RGB and RGB-D based meth ods for mirror segmentation. Experimental results also show that depth is a powe rful cue for mirror segmentation.
****
You Only Look One-Level Feature
Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, Jian Sun; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2021, pp. 13039-13048
This paper revisits feature pyramids networks (FPN) for one-stage detectors and points out that the success of FPN is due to its divide-and-conquer solution to the optimization problem in object detection rather than multi-scale feature fus ion. From the perspective of optimization, we introduce an alternative way to ad dress the problem instead of adopting the complex feature pyramids -- utilizing only one-level feature for detection. Based on the simple and efficient solution , we present You Only Look One-level Feature (YOLOF). In our method, two key com

ponents, Dilated Encoder and Uniform Matching, are proposed and bring considerab le improvements. Extensive experiments on the COCO benchmark prove the effective ness of the proposed model. Our YOLOF achieves comparable results with its featu re pyramids counterpart RetinaNet while being 2.5 times faster. Without transfor mer layers, YOLOF can match the performance of DETR in a single-level feature ma nner with 7 times less training epochs.
****
Multi-Perspective LSTM for Joint Visual Representation Learning
Alireza Sepas-Moghaddam, Fernando Pereira, Paulo Lobato Correia, Ali Etemad; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16540-16548
We present a novel LSTM cell architecture capable of learning both intra- and in ter-perspective relationships available in visual sequences captured from multip le perspectives. Our architecture adopts a novel recurrent joint learning strate gy that uses additional gates and memories at the cell level. We demonstrate tha t by using the proposed cell to create a network, more effective and richer visu al representations are learned for recognition tasks. We validate the performanc e of our proposed architecture in the context of two multi-perspective visual re cognition tasks namely lip reading and face recognition. Three relevant datasets are considered and the results are compared against fusion strategies, other ex isting multi-input LSTM architectures, and alternative recognition solutions. Th e experiments show the superior performance of our solution over the considered benchmarks, both in terms of recognition accuracy and complexity. We make our co de publicly available at: https://github.com/arsm/MPLSTM
****
Towards Improving the Consistency, Efficiency, and Flexibility of Differentiable Neural Architecture Search
Yibo Yang, Shan You, Hongyang Li, Fei Wang, Chen Qian, Zhouchen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 021, pp. 6667-6676
Most differentiable neural architecture search methods construct a super-net for search and derive a target-net as its sub-graph for evaluation. There exists a significant gap between the architectures in search and evaluation. As a result, current methods suffer from an inconsistent, inefficient, and inflexible search process. In this paper, we introduce EnTranNAS that is composed of Engine-cells and Transit-cells. The Engine-cell is differentiable for architecture search, w hile the Transit-cell only transits a sub-graph by architecture derivation. Cons equently, the gap between the architectures in search and evaluation is signific antly reduced. Our method also spares much memory and computation cost, which sp eeds up the search process. A feature sharing strategy is introduced for more ba lanced optimization and more efficient search. Furthermore, we develop an archit ecture derivation method to replace the traditional one that is based on a hand- crafted rule. Our method enables differentiable sparsification, and keeps the de rived architecture equivalent to that of Engine-cell, which further improves the consistency between search and evaluation. More importantly, it supports the se arch for topology where a node can be connected to prior nodes with any number o f connections, so that the searched architectures could be more flexible. Our se arch on CIFAR-10 has an error rate of 2.22% with only 0.07 GPU-day. We can also directly perform the search on ImageNet with topology learnable and achieve a to p-1 error rate of 23.8% in 2.1 GPU-day.
****
Gaussian Context Transformer
Dongsheng Ruan, Daiyin Wang, Yuan Zheng, Nenggan Zheng, Min Zheng; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 1, pp. 15129-15138
Recently, a large number of channel attention blocks are proposed to boost the r epresentational power of deep convolutional neural networks (CNNs). These approa ches commonly learn the relationship between global contexts and attention activ ations by using fully-connected layers or linear transformations. However, we em pirically find that though many parameters are introduced, these attention block

s may not learn the relationship well. In this paper, we hypothesize that the relationship is predetermined. Based on this hypothesis, we propose a simple yet extremely efficient channel attention block, called Gaussian Context Transformer (GCT), which achieves contextual feature excitation using a Gaussian function that satisfies the presupposed relationship. According to whether the standard deviation of the Gaussian function is learnable, we develop two versions of GCT: GCT-B0 and GCT-B1. GCT-B0 is a parameter-free channel attention block by fixing the standard deviation. It directly maps global contexts to attention activations without learning. In contrast, GCT-B1 is a parameterized channel attention block, which adaptively learns the standard deviation to enhance the mapping ability. Extensive experiments on ImageNet and MS COCO benchmarks demonstrate that our GCTs lead to consistent improvements across various deep CNNs and detectors. Compared with a bank of state-of-the-art channel attention blocks, such as SE and ECA , our GCTs are superior in effectiveness and efficiency.
****

Keypoint-Graph-Driven Learning Framework for Object Pose Estimation
Shaobo Zhang, Wanqing Zhao, Ziyu Guan, Xianlin Peng, Jinye Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1065-1073
Many recent 6D pose estimation methods exploited object 3D models to generate synthetic images for training because labels come for free. However, due to the domain shift of data distributions between real images and synthetic images, the network trained only on synthetic images fails to capture robust features in real images for 6D pose estimation. We propose to solve this problem by making the network insensitive to different domains, rather than taking the more difficult route of forcing synthetic images to be similar to real images. Inspired by domain adaption methods, a Domain Adaptive Keypoints Detection Network (DAKDN) including a domain adaption layer is used to minimize the discrepancy of deep features between synthetic and real images. A unique challenge here is the lack of ground truth labels (i.e., keypoints) for real images. Fortunately, the geometry relations between keypoints are invariant under real/synthetic domains. Hence, we propose to use the domain-invariant geometry structure among keypoints as a "bridge" constraint to optimize DAKDN for 6D pose estimation across domains. Specifically, DAKDN employs a Graph Convolutional Network (GCN) block to learn the geometry structure from synthetic images and uses the GCN to guide the training for real images. The 6D poses of objects are calculated using Perspective-n-Point (PnP) algorithm based on the predicted keypoints. Experiments show that our method outperforms state-of-the-art approaches without manual poses labels and competes with approaches using manual poses labels.
****

Deep Burst Super-Resolution
Goutam Bhat, Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9209-9218
While single-image super-resolution (SISR) has attracted substantial interest in recent years, the proposed approaches are limited to learning image priors in order to add high frequency details. In contrast, multi-frame super-resolution (MFSR) offers the possibility of reconstructing rich details by combining signal information from multiple shifted images. This key advantage, along with the increasing popularity of burst photography, have made MFSR an important problem for real-world applications. We propose a novel architecture for the burst super-resolution task. Our network takes multiple noisy RAW images as input, and generates a denoised, super-resolved RGB image as output. This is achieved by explicitly aligning deep embeddings of the input frames using pixel-wise optical flow. The information from all frames are then adaptively merged using an attention-based fusion module. In order to enable training and evaluation on real-world data, we additionally introduce the BurstSR dataset, consisting of smartphone bursts and high-resolution DSLR ground-truth. We perform comprehensive experimental analysis, demonstrating the effectiveness of the proposed architecture.
****

Transferable Semantic Augmentation for Domain Adaptation
Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Wei Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11516-11525
Domain adaptation has been widely explored by transferring the knowledge from a label-rich source domain to a related but unlabeled target domain. Most existing domain adaptation algorithms attend to adapting feature representations across two domains with the guidance of a shared source-supervised classifier. However, such classifier limits the generalization ability towards unlabeled target recognition. To remedy this, we propose a Transferable Semantic Augmentation (TSA) approach to enhance the classifier adaptation ability through implicitly generating source features towards target semantics. Specifically, TSA is inspired by the fact that deep feature transformation towards a certain direction can be represented as meaningful semantic altering in the original input space. Thus, source features can be augmented to effectively equip with target semantics to train a more transferable classifier. To achieve this, for each class, we first use the inter-domain feature mean difference and target intra-class feature covariance to construct a multivariate normal distribution. Then we augment source features with random directions sampled from the distribution class-wisely. Interestingly, such source augmentation is implicitly implemented through an expected transferable cross-entropy loss over the augmented source distribution, where an upper bound of the expected loss is derived and minimized, introducing negligible computational overhead. As a light-weight and general technique, TSA can be easily plugged into various domain adaptation methods, bringing remarkable improvements. Comprehensive experiments on cross-domain benchmarks validate the efficacy of TSA.
****
Patchwise Generative ConvNet: Training Energy-Based Models From a Single Natural Image for Internal Learning
Zilong Zheng, Jianwen Xie, Ping Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2961-2970
Exploiting internal statistics of a single natural image has long been recognized as a significant research paradigm where the goal is to learn the distribution of patches within the image without relying on external training data. Different from prior works that model such distributions implicitly with a top-down latent variable model (i.e., generator), in this work, we propose to explicitly represent the statistical distribution within a single natural image by using an energy-based generative framework, where a pyramid of energy functions parameterized by bottom-up deep neural networks, are used to capture the distributions of patches at different resolutions. Meanwhile, a coarse-to-fine sequential training and sampling strategy is presented to train the model efficiently. Besides learning to generate random samples from white noise, the model can learn in parallel to recover a real image from its incomplete version, which can improve the descriptive power of the learned models. The proposed model not only is simple and natural in that it does not require auxiliary models (e.g., discriminators) to assist the training, but also unifies internal statistics learning and image generation in a single framework. Qualitative results are presented on various image generation tasks, including super-resolution, image editing, harmonization, etc. The evaluation and user studies demonstrate the superior quality of our results.
****
Clusformer: A Transformer Based Clustering Approach to Unsupervised Large-Scale Face and Visual Landmark Recognition
Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D. Bui, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10847-10856
The research in automatic unsupervised visual clustering has received considerable attention over the last couple years. It aims at explaining distributions of unlabeled visual images by clustering them via a parameterized model of appearance. Graph Convolutional Neural Networks (GCN) have recently been one of the most

popular clustering methods. However, it has reached some limitations. Firstly, it is quite sensitive to hard or noisy samples. Secondly, it is hard to investigate with various deep network models due to its computational training time. Finally, it is hard to design an end-to-end training model between the deep feature extraction and GCN clustering modeling. This work therefore presents the Clusformer, a simple but new perspective of Transformer based approach, to automatic visual clustering via its unsupervised attention mechanism. The proposed method is able to robustly deal with noisy or hard samples. It is also flexible and effective to collaborate with different deep network models with various model sizes in an end-to-end framework. The proposed method is evaluated on two popular large-scale visual databases, i.e. Google Landmark and MS-Celeb-1M face database, and outperforms prior unsupervised clustering methods. Code will be available at https://github.com/VinAIResearch/Clusformer
****
No Frame Left Behind: Full Video Action Recognition
Xin Liu, Silvia L. Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, Jan C. van Gemert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14892-14901
Not all video frames are equally informative for recognizing an action. It is computationally infeasible to train deep networks on all video frames when actions develop over hundreds of frames. A common heuristic is uniformly sampling a small number of video frames and using these to recognize the action. Instead, here we propose full video action recognition and consider all video frames. To make this computational tractable, we first cluster all frame activations along the temporal dimension based on their similarity with respect to the classification task, and then temporally aggregate the frames in the clusters into a smaller number of representations. Our method is end-to-end trainable and computationally efficient as it relies on temporally localized clustering in combination with fast Hamming distances in feature space. We evaluate on UCF101, HMDB51, Breakfast, and Something-Something V1 and V2, where we compare favorably to existing heuristic frame sampling methods.
****
ColorRL: Reinforced Coloring for End-to-End Instance Segmentation
Tran Anh Tuan, Nguyen Tuan Khoa, Tran Minh Quan, Won-Ki Jeong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16727-16736
Instance segmentation, the task of identifying and separating each individual object of interest in the image, is one of the actively studied research topics in computer vision. Although many feed-forward networks produce high-quality binary segmentation on different types of images, their final result heavily relies on the post-processing step, which separates instances from the binary mask. In comparison, the existing iterative methods extract a single object at a time using discriminative knowledge-based properties (e.g., shapes, boundaries, etc.) without relying on post-processing. However, they do not scale well with a large number of objects. To exploit the advantages of conventional sequential segmentation methods without impairing the scalability, we propose a novel iterative deep reinforcement learning agent that learns how to differentiate multiple objects in parallel. By constructing a relational graph between pixels, we design a reward function that encourages separating pixels of different objects and grouping pixels that belong to the same instance. We demonstrate that the proposed method can efficiently perform instance segmentation of many objects without heavy post-processing.
****
Compatibility-Aware Heterogeneous Visual Search
Rahul Duggal, Hao Zhou, Shuo Yang, Yuanjun Xiong, Wei Xia, Zhuowen Tu, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10723-10732
We tackle the problem of visual search under resource constraints. Existing systems use the same embedding model to compute representations (embeddings) for the query and gallery images. Such systems inherently face a hard accuracy-efficien

cy trade-off: the embedding model needs to be large enough to ensure high accuracy, yet small enough to enable query-embedding computation on resource-constrained platforms. This trade-off could be mitigated if gallery embeddings are generated from a large model and query embeddings are extracted using a compact model. The key to building such a system is to ensure representation compatibility between the query and gallery models. In this paper, we address two forms of compatibility: One enforced by modifying the parameters of each model that computes the embeddings. The other by modifying the architectures that compute the embeddings, leading to compatibility-aware neural architecture search (CMP-NAS). We test CMP-NAS on challenging retrieval tasks for fashion images (DeepFashion2), and face images (IJB-C). Compared to ordinary (homogeneous) visual search using the largest embedding model (paragon), CMP-NAS achieves 80-fold and 23-fold cost reduction while maintaining accuracy within 0.3% and 1.6% of the paragon on DeepFashion2 and IJB-C respectively.
****
WOAD: Weakly Supervised Online Action Detection in Untrimmed Videos
Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, Caiming Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1915-1923
Online action detection in untrimmed videos aims to identify an action as it happens, which makes it very important for real-time applications. Previous methods rely on tedious annotations of temporal action boundaries for training, which hinders the scalability of online action detection systems. We propose WOAD, a weakly supervised framework that can be trained using only video-class labels. WOAD contains two jointly-trained modules, i.e., temporal proposal generator (TPG) and online action recognizer (OAR). Supervised by video-class labels, TPG works offline and targets at accurately mining pseudo frame-level labels for OAR. With the supervisory signals from TPG, OAR learns to conduct action detection in an online fashion. Experimental results on THUMOS'14, ActivityNet1.2 and ActivityNet1.3 show that our weakly-supervised method largely outperforms weakly-supervised baselines and achieves comparable performance to the previous strongly-supervised methods. Beyond that, WOAD is flexible to leverage strong supervision when it is available. When strongly supervised, our method obtains the state-of-the-art results in the tasks of both online per-frame action recognition and online detection of action start.
****
Deep Dual Consecutive Network for Human Pose Estimation
Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, Xun Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 525-534
Multi-frame human pose estimation in complicated situations is challenging. Although state-of-the-art human joints detectors have demonstrated remarkable results for static images, their performances come short when we apply these models to video sequences. Prevalent shortcomings include the failure to handle motion blur, video defocus, or pose occlusions, arising from the inability in capturing the temporal dependency among video frames. On the other hand, directly employing conventional recurrent neural networks incurs empirical difficulties in modeling spatial contexts, especially for dealing with pose occlusions. In this paper, we propose a novel multi-frame human pose estimation framework, leveraging abundant temporal cues between video frames to facilitate keypoint detection. Three modular components are designed in our framework. A Pose Temporal Merger encodes keypoint spatiotemporal context to generate effective searching scopes while a Pose Residual Fusion module computes weighted pose residuals in dual directions. These are then processed via our Pose Correction Network for efficient refining of pose estimations. Our method ranks No.1 in the Multi-frame Person Pose Estimation Challenge on the large-scale benchmark datasets PoseTrack2017 and PoseTrack2018. We have released our code, hoping to inspire future research.
****
Uncertainty-Aware Joint Salient Object and Camouflaged Object Detection
Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, Yuchao Dai; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10071-10081
Visual salient object detection (SOD) aims at finding the salient object(s) that attract human attention, while camouflaged object detection (COD) on the contrary intends to discover the camouflaged object(s) that hidden in the surrounding. In this paper, we propose a paradigm of leveraging the contradictory information to enhance the detection ability of both salient object detection and camouflaged object detection. We start by exploiting the easy positive samples in the COD dataset to serve as hard positive samples in the SOD task to improve the robustness of the SOD model. Then, we introduce a \enquote similarity measure module to explicitly model the contradicting attributes of these two tasks. Furthermore, considering the uncertainty of labeling in both tasks' datasets, we propose an adversarial learning network to achieve both higher order similarity measure and network confidence estimation. Experimental results on benchmark datasets demonstrate that our solution leads to state-of-the-art (SOTA) performance for both tasks.
****
HourNAS: Extremely Fast Neural Architecture Search Through an Hourglass Lens
Zhaohui Yang, Yunhe Wang, Xinghao Chen, Jianyuan Guo, Wei Zhang, Chao Xu, Chunjing Xu, Dacheng Tao, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10896-10906
Neural Architecture Search (NAS) aims to automatically discover optimal architectures. In this paper, we propose an hourglass-inspired approach (HourNAS) for extremely fast NAS. It is motivated by the fact that the effects of the architecture often proceed from the vital few blocks. Acting like the narrow neck of an hourglass, vital blocks in the guaranteed path from the input to the output of a deep neural network restrict the information flow and influence the network accuracy. The other blocks occupy the major volume of the network and determine the overall network complexity, corresponding to the bulbs of an hourglass. To achieve an extremely fast NAS while preserving the high accuracy, we propose to identify the vital blocks and make them the priority in the architecture search. The search space of those non-vital blocks is further shrunk to only cover the candidates that are affordable under the computational resource constraints. Experimental results on ImageNet show that only using 3 hours (0.1 days) with one GPU, our HourNAS can search an architecture that achieves a 77.0% Top-1 accuracy, which outperforms the state-of-the-art methods.
****
Tree-Like Decision Distillation
Jie Song, Haofei Zhang, Xinchao Wang, Mengqi Xue, Ying Chen, Li Sun, Dacheng Tao, Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13488-13497
Knowledge distillation pursues a diminutive yet well-behaved student network by harnessing the knowledge learned by a cumbersome teacher model. Prior methods achieve this by making the student imitate shallow behaviors, such as soft targets, features, or attention, of the teacher. In this paper, we argue that what really matters for distillation is the intrinsic problem-solving process captured by the teacher. By dissecting the decision process in a layer-wise manner, we found that the decision-making procedure in the teacher model is conducted in a coarse-to-fine manner, where coarse-grained discrimination (e.g., animal vs vehicle) is attained in early layers, and fine-grained discrimination (e.g., dog vs cat, car vs truck) in latter layers. Motivated by this observation, we propose a new distillation method, dubbed as Tree-like Decision Distillation (TDD), to endow the student with the same problem-solving mechanism as that of the teacher. Extensive experiments demonstrated that TDD yields competitive performance compared to state of the arts. More importantly, it enjoys better interpretability due to its interpretable decision distillation instead of dark knowledge distillation.
****
GAN Prior Embedded Network for Blind Face Restoration in the Wild
Tao Yang, Peiran Ren, Xuansong Xie, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 672-681

Blind face restoration (BFR) from severely degraded face images in the wild is a very challenging problem. Due to the high illness of the problem and the complex unknown degradation, directly training a deep neural network (DNN) usually cannot lead to acceptable results. Existing generative adversarial network (GAN) based methods can produce better results but tend to generate over-smoothed restorations. In this work, we propose a new method by first learning a GAN for high-quality face image generation and embedding it into a U-shaped DNN as a prior decoder, then fine-tuning the GAN prior embedded DNN with a set of synthesized low-quality face images. The GAN blocks are designed to ensure that the latent code and noise input to the GAN can be respectively generated from the deep and shallow features of the DNN, controlling the global face structure, local face details and background of the reconstructed image. The proposed GAN prior embedded network (GPEN) is easy-to-implement, and it can generate visually photo-realistic results. Our experiments demonstrated that the proposed GPEN achieves significantly superior results to state-of-the-art BFR methods both quantitatively and qualitatively, especially for the restoration of severely degraded face images in the wild. The source code and models can be found at https://github.com/yangxy/GPEN.
****
Collaborative Spatial-Temporal Modeling for Language-Queried Video Actor Segmentation
Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, Fei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4187-4196
Language-queried video actor segmentation aims to predict the pixel-level mask of the actor which performs the actions described by a natural language query in the target frames. Existing methods adopt 3D CNNs over the video clip as a general encoder to extract a mixed spatio-temporal feature for the target frame. Though 3D convolutions are amenable to recognizing which actor is performing the queried actions, it also inevitably introduces misaligned spatial information from adjacent frames, which confuses features of the target frame and yields inaccurate segmentation. Therefore, we propose a collaborative spatial-temporal encoder-decoder framework which contains a 3D temporal encoder over the video clip to recognize the queried actions, and a 2D spatial encoder over the target frame to accurately segment the queried actors. In the decoder, a Language-Guided Feature Selection (LGFS) module is proposed to flexibly integrate spatial and temporal features from the two encoders. We also propose a Cross-Modal Adaptive Modulation (CMAM) module to dynamically recombine spatial- and temporal-relevant linguistic features for multimodal feature interaction in each stage of the two encoders. Our method achieves new state-of-the-art performance on two popular benchmarks with less computational overhead than previous approaches.
****
Drafting and Revision: Laplacian Pyramid Network for Fast High-Quality Artistic Style Transfer
Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, Xinbo Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5141-5150
Artistic style transfer aims at migrating the style from an example image to a content image. Currently, optimization-based methods have achieved great stylization quality, but expensive time cost restricts their practical applications. Meanwhile, feed-forward methods still fail to synthesize complex style, especially when holistic global and local patterns exist. Inspired by the common painting process of drawing a draft and revising the details, we introduce a novel feed-forward method Laplacian Pyramid Network (LapStyle). LapStyle first transfers global style pattern in low-resolution via a Drafting Network. It then revises the local details in high-resolution via a Revision Network, which hallucinates a residual image according to the draft and the image textures extracted by Laplacian filtering. Higher resolution details can be easily generated by stacking Revision Networks with multiple Laplacian pyramid levels. The final stylized image is obtained by aggregating outputs of all pyramid levels. Experiments demonstrate t

hat our method can synthesize high quality stylized images in real time, where holistic style patterns are properly transferred.
****
The Lottery Ticket Hypothesis for Object Recognition
Sharath Girish, Shishira R Maiya, Kamal Gupta, Hao Chen, Larry S. Davis, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 762-771
Recognition tasks, such as object recognition and keypoint estimation, have seen widespread adoption in recent years. Most state-of-the-art methods for these tasks use deep networks that are computationally expensive and have huge memory footprints. This makes it exceedingly difficult to deploy these systems on low power embedded devices. Hence, the importance of decreasing the storage requirements and the amount of computation in such models is paramount. The recently proposed Lottery Ticket Hypothesis (LTH) states that deep neural networks trained on large datasets contain smaller subnetworks that achieve on par performance as the dense networks. In this work, we perform the first empirical study investigating LTH for model pruning in the context of object detection, instance segmentation, and keypoint estimation. Our studies reveal that lottery tickets obtained from ImageNet pretraining do not transfer well to the downstream tasks. We provide guidance on how to find lottery tickets with up to 80% overall sparsity on different sub-tasks without incurring any drop in the performance. Finally, we analyse the behavior of trained tickets with respect to various task attributes such as object size, frequency, and difficulty of detection.
****
Refer-It-in-RGBD: A Bottom-Up Approach for 3D Visual Grounding in RGBD Images
Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, Shuguang Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6032-6041
Grounding referring expressions in RGBD image has been an emerging field. We present a novel task of 3D visual grounding in single-view RGBD image where the referred objects are often only partially scanned due to occlusion. In contrast to previous works that directly generate object proposals for grounding in the 3D scenes, we propose a bottom-up approach to gradually aggregate content-aware information, effectively addressing the challenge posed by the partial geometry. Our approach first fuses the language and the visual features at the bottom level to generate a heatmap that coarsely localizes the relevant regions in the RGBD image. Then our approach conducts an adaptive feature learning based on the heatmap and performs the object-level matching with another visio-linguistic fusion to finally ground the referred object. We evaluate the proposed method by comparing to the state-of-the-art methods on both the RGBD images extracted from the ScanRefer dataset and our newly collected SUNRefer dataset. Experiments show that our method outperforms the previous methods by a large margin (by 11.2% and 15.6% Acc@0.5) on both datasets.
****
LQF: Linear Quadratic Fine-Tuning
Alessandro Achille, Aditya Golatkar, Avinash Ravichandran, Marzia Polito, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15729-15739
Classifiers that are linear in their parameters, and trained by optimizing a convex loss function, have predictable behavior with respect to changes in the training data, initial conditions, and optimization. Such desirable properties are absent in deep neural networks (DNNs), typically trained by non-linear fine-tuning of a pre-trained model. Previous attempts to linearize DNNs have led to interesting theoretical insights, but have not impacted the practice due to the substantial performance gap compared to standard non-linear optimization. We present the first method for linearizing a pre-trained model that achieves comparable performance to non-linear fine-tuning on most of real-world image classification tasks tested, thus enjoying the interpretability of linear models without incurring punishing losses in performance. LQF consists of simple modifications to the architecture, loss function and optimization typically used for classification: L

eaky-ReLU instead of ReLU, mean squared loss instead of cross-entropy, and pre-conditioning using Kronecker factorization. None of these changes in isolation is sufficient to approach the performance of non-linear fine-tuning. When used in combination, they allow us to reach comparable performance, and even superior in the low-data regime, while enjoying the simplicity, robustness and interpretability of linear-quadratic optimization.
****
Watching You: Global-Guided Reciprocal Learning for Video-Based Person Re-Identification
Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, Xiaoyun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13334-13343
Video-based person re-identification (Re-ID) aims to automatically retrieve video sequences of the same person under non-overlapping cameras. To achieve this goal, it is the key to fully utilize abundant spatial and temporal cues in videos. Existing methods usually focus on the most conspicuous image regions, thus they may easily miss out fine-grained clues due to the person varieties in image sequences. To address above issues, in this paper, we propose a novel Global-guided Reciprocal Learning (GRL) framework for video-based person Re-ID. Specifically, we first propose a Global-guided Correlation Estimation (GCE) to generate feature correlation maps of local features and global features, which help to localize the high- and low-correlation regions for identifying the same person. After that, the discriminative features are disentangled into high-correlation features and low-correlation features under the guidance of the global representations. Moreover, a novel Temporal Reciprocal Learning (TRL) mechanism is designed to sequentially enhance the high-correlation semantic information and accumulate the low-correlation sub-critical clues. Extensive experiments are conducted on three public benchmarks. The experimental results indicate that our approach can achieve better performance than other state-of-the-art approaches. The code is released at https://github.com/flysnowtiger/GRL.
****
S3: Learnable Sparse Signal Superdensity for Guided Depth Estimation
Yu-Kai Huang, Yueh-Cheng Liu, Tsung-Han Wu, Hung-Ting Su, Yu-Cheng Chang, Tsung-Lin Tsou, Yu-An Wang, Winston H. Hsu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16706-16716
Dense depth estimation plays a key role in multiple applications such as robotics, 3D reconstruction, and augmented reality. While sparse signal, e.g., LiDAR and Radar, has been leveraged as guidance for enhancing dense depth estimation, the improvement is limited due to its low density and imbalanced distribution. To maximize the utility from the sparse source, we propose Sparse Signal Superdensity (S3) technique, which expands the depth value from sparse cues while estimating the confidence of expanded region. The proposed S3 can be applied to various guided depth estimation approaches and trained end-to-end at different stages, including input, cost volume and output. Extensive experiments demonstrate the effectiveness, robustness, and flexibility of the S3 technique on LiDAR and Radar signal.
****
Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking
Ning Wang, Wengang Zhou, Jie Wang, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1571-1580
In video object tracking, there exist rich temporal contexts among successive frames, which have been largely overlooked in existing trackers. In this work, we bridge the individual video frames and explore the temporal contexts across them via a transformer architecture for robust object tracking. Different from classic usage of the transformer in natural language processing tasks, we separate its encoder and decoder into two parallel branches and carefully design them within the Siamese-like tracking pipelines. The transformer encoder promotes the target templates via attention-based feature reinforcement, which benefits the high-quality tracking model generation. The transformer decoder propagates the tracki

ng cues from previous templates to the current frame, which facilitates the obje ct searching process. Our transformer-assisted tracking framework is neat and tr ained in an end-to-end manner. With the proposed transformer, a simple Siamese m atching approach is able to outperform the current top-performing trackers. By c ombining our transformer with the recent discriminative tracking pipeline, our m ethod sets several new state-of-the-art records on prevalent tracking benchmarks .
****
High-Fidelity Neural Human Motion Transfer From Monocular Video
Moritz Kappel, Vladislav Golyanik, Mohamed Elgharib, Jann-Ole Henningson, Hans-P eter Seidel, Susana Castillo, Christian Theobalt, Marcus Magnor; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1541-1550
Video-based human motion transfer creates video animations of humans following a source motion. Current methods show remarkable results for tightly-clad subject s. However, the lack of temporally consistent handling of plausible clothing dyn amics, including fine and high-frequency details, significantly limits the attai nable visual quality. We address these limitations for the first time in the lit erature and present a new framework which performs high-fidelity and temporally- consistent human motion transfer with natural pose-dependent non-rigid deformati ons, for several types of loose garments. In contrast to the previous techniques , we perform image generation in three subsequent stages: synthesizing human sha pe, structure, and appearance. Given a monocular RGB video of an actor, we train a stack of recurrent deep neural networks that generate these intermediate repr esentations from 2D poses and their temporal derivatives. Splitting the difficul t motion transfer problem into subtasks that are aware of the temporal motion co ntext helps us to synthesize results with plausible dynamics and pose-dependent detail. It also allows artistic control of results by manipulation of individual framework stages. In the experimental results, we significantly outperform the state-of-the-art in terms of video realism. The source code is available at http s://graphics.tu-bs.de/publications/kappel2020high-fidelity.
****
Polygonal Building Extraction by Frame Field Learning
Nicolas Girard, Dmitriy Smirnov, Justin Solomon, Yuliya Tarabalka; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 1, pp. 5891-5900
While state of the art image segmentation models typically output segmentations in raster format, applications in geographic information systems often require v ector polygons. To help bridge the gap between deep network output and the forma t used in downstream tasks, we add a frame field output to a deep segmentation m odel for extracting buildings from remote sensing images. We train a deep neural network that aligns a predicted frame field to ground truth contours. This addi tional objective improves segmentation quality by leveraging multi-task learning and provides structural information that later facilitates polygonization; we a lso introduce a polygonization algorithm that that utilizes the frame field alon g with the raster segmentation. Our code is available at https://github.com/Lydo rn/Polygonization-by-Frame-Field-Learning.
****
NeuralFusion: Online Depth Fusion in Latent Space
Silvan Weder, Johannes L. Schonberger, Marc Pollefeys, Martin R. Oswald; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3162-3172
We present a novel online depth map fusion approach that learns depth map aggreg ation in a latent feature space. While previous fusion methods use an explicit s cene representation like signed distance functions (SDFs), we propose a learned feature representation for the fusion. The key idea is a separation between the scene representation used for the fusion and the output scene representation, vi a an additional translator network. Our neural network architecture consists of two main parts: a depth and feature fusion sub-network, which is followed by a t ranslator sub-network to produce the final surface representation (e.g. TSDF) fo

r visualization or other tasks. Our approach is an online process, handles high noise levels, and is particularly able to deal with gross outliers common for photometric stereo-based depth maps. Experiments on real and synthetic data demonstrate improved results compared to the state of the art, especially in challenging scenarios with large amounts of noise and outliers. The source code will be made available at https://github.com/weders/NeuralFusion.
****
PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation
Kehong Gong, Jianfeng Zhang, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8575-8584
Existing 3D human pose estimators suffer poor generalization performance to new datasets, largely due to the limited diversity of 2D-3D pose pairs in the training data. To address this problem, we present PoseAug, a new auto-augmentation framework that learns to augment the available training poses towards a greater diversity and thus improve generalization of the trained 2D-to-3D pose estimator. Specifically, PoseAug introduces a novel pose augmentor that learns to adjust various geometry factors (e.g., posture, body size, view point and position) of a pose through differentiable operations. With such differentiable capacity, the augmentor can be jointly optimized with the 3D pose estimator and take the estimation error as feedback to generate more diverse and harder poses in an online manner. Moreover, PoseAug introduces a novel part-aware Kinematic Chain Space for evaluating local joint-angle plausibility and develops a discriminative module accordingly to ensure the plausibility of the augmented poses. These elaborate designs enable PoseAug to generate more diverse yet plausible poses than existing offline augmentation methods, and thus yield better generalization of the pose estimator. PoseAug is generic and easy to be applied to various 3D pose estimators. Extensive experiments demonstrate that PoseAug brings clear improvements on both intra-scenario and cross-scenario datasets. Notably, it achieves 88.6% 3D PCK on MPI-INF-3DHP under cross-dataset evaluation setup, improving upon the previous best data augmentation based method by 9.1%. Code can be found at: https://github.com/jfzhang95/PoseAug.
****
Depth Completion With Twin Surface Extrapolation at Occlusion Boundaries
Saif Imran, Xiaoming Liu, Daniel Morris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2583-2592
Depth completion starts from a sparse set of known depth values and estimates the unknown depths for the remaining image pixels. Most methods model this as depth interpolation and erroneously interpolate depth pixels into the empty space between spatially distinct objects, resulting in depth-smearing across occlusion boundaries. Here we propose a multi-hypothesis depth representation that explicitly models both foreground and background depths in the difficult occlusion-boundary regions. Our method can be thought of as performing twin-surface extrapolation, rather than interpolation, in these regions. Next our method fuses these extrapolated surfaces into a single depth image leveraging the image data. Key to our method is the use of an asymmetric loss function that operates on a novel twin-surface representation. This enables us to train a network to simultaneously do surface extrapolation and surface fusion. We characterize our loss function and compare with other common losses. Finally, we validate our method on three different datasets; KITTI, an outdoor real-world dataset, NYU2, indoor real-world depth dataset and Virtual KITTI, a photo-realistic synthetic dataset with dense groundtruth, and demonstrate improvement over the state of the art.
****
Learning the Superpixel in a Non-Iterative and Lifelong Manner
Lei Zhu, Qi She, Bin Zhang, Yanye Lu, Zhilin Lu, Duo Li, Jie Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1225-1234
Superpixel is generated by automatically clustering pixels in an image into hundreds of compact partitions, which is widely used to perceive the object contours for its excellent contour adherence. Although some works use the Convolution Ne

ural Network (CNN) to generate high-quality superpixel, we challenge the design principles of these networks, specifically for their dependence on manual labels and excess computation resources, which limits their flexibility compared with the traditional unsupervised segmentation methods. We target at redefining the CNN-based superpixel segmentation as a lifelong clustering task and propose an unsupervised CNN-based method called LNS-Net. The LNS-Net can learn superpixel in a non-iterative and lifelong manner without any manual labels. Specifically, a lightweight feature embedder is proposed for LNS-Net to efficiently generate the cluster-friendly features. With those features, seed nodes can be automatically assigned to cluster pixels in a non-iterative way. Additionally, our LNS-Net can adapt the sequentially lifelong learning by rescaling the gradient of weight based on both channel and spatial context to avoid overfitting. Experiments show that the proposed LNS-Net achieves significantly better performance on three benchmarks with nearly ten times lower complexity compared with other state-of-the-art methods.
****
Image Generators With Conditionally-Independent Pixel Synthesis
Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, Denis Korzhenkov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14278-14287
Existing image generator networks rely heavily on spatial convolutions and, optionally, self-attention blocks in order to gradually synthesize images in a coarse-to-fine manner. Here, we present a new architecture for image generators, where the color value at each pixel is computed independently given the value of a random latent vector and the coordinate of that pixel. No spatial convolutions or similar operations that propagate information across pixels are involved during the synthesis. We analyze the modeling capabilities of such generators when trained in an adversarial fashion, and observe the new generators to achieve similar generation quality to state-of-the-art convolutional generators. We also investigate several interesting properties unique to the new architecture.
****
Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets
Yuan-Hong Liao, Amlan Kar, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4350-4359
Data is the engine of modern computer vision, which necessitates collecting large-scale datasets. This is expensive, and guaranteeing the quality of the labels is a major challenge. In this paper, we investigate efficient annotation strategies for collecting multi-class classification labels for a large collection of images. While methods that exploit learnt models for labeling exist, a surprisingly prevalent approach is to query humans for a fixed number of labels per datum and aggregate them, which is expensive. Building on prior work on online joint probabilistic modeling of human annotations and machine-generated beliefs, we propose modifications and best practices aimed at minimizing human labeling effort. Specifically, we make use of advances in self-supervised learning, view annotation as a semi-supervised learning problem, identify and mitigate pitfalls and ablate several key design choices to propose effective guidelines for labeling. Our analysis is done in a more realistic simulation that involves querying human labelers, which uncovers issues with evaluation using existing worker simulation methods. Simulated experiments on a 125k image subset of the ImageNet100 show that it can be annotated to 80% top-1 accuracy with 0.35 annotations per image on average, a 2.7x and 6.7x improvement over prior work and manual annotation, respectively.
****
Seesaw Loss for Long-Tailed Instance Segmentation
Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9695-9704
Instance segmentation has witnessed a remarkable progress on class-balanced benchmarks. However, they fail to perform as accurately in real-world scenarios, whe

re the category distribution of objects naturally comes with a long tail. Instances of head classes dominate a long-tailed dataset and they serve as negative samples of tail categories. The overwhelming gradients of negative samples on tail classes lead to a biased learning process for classifiers. Consequently, objects of tail categories are more likely to be misclassified as backgrounds or head categories. To tackle this problem, we propose Seesaw Loss to dynamically re-balance gradients of positive and negative samples for each category, with two complementary factors, i.e., mitigation factor and compensation factor. The mitigation factor reduces punishments to tail categories w.r.t. the ratio of cumulative training instances between different categories. Meanwhile, the compensation factor increases the penalty of misclassified instances to avoid false positives of tail categories. We conduct extensive experiments on Seesaw Loss with mainstream frameworks and different data sampling strategies. With a simple end-to-end training pipeline, Seesaw Loss obtains significant gains over Cross-Entropy Loss, and achieves state-of-the-art performance on LVIS dataset without bells and whistles. Code is available at https://github.com/open-mmlab/mmdetection.
****
Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction
Guy Gafni, Justus Thies, Michael Zollhofer, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8649-8658
We present dynamic neural radiance fields for modeling the appearance and dynamics of a human face. Digitally modeling and reconstructing a talking human is a key building-block for a variety of applications. Especially, for telepresence applications in AR or VR, a faithful reproduction of the appearance including novel viewpoint or head-poses is required. In contrast to state-of-the-art approaches that model the geometry and material properties explicitly, or are purely image-based, we introduce an implicit representation of the head based on scene representation networks. To handle the dynamics of the face, we combine our scene representation network with a low-dimensional morphable model which provides explicit control over pose and expressions. We use volumetric rendering to generate images from this hybrid representation and demonstrate that such a dynamic neural scene representation can be learned from monocular input data only, without the need of a specialized capture setup. In our experiments, we show that this learned volumetric representation allows for photorealistic image generation that surpasses the quality of state-of-the-art video-based reenactment methods.
****
PU-GCN: Point Cloud Upsampling Using Graph Convolutional Networks
Guocheng Qian, Abdulellah Abualshour, Guohao Li, Ali Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11683-11692
The effectiveness of learning-based point cloud upsampling pipelines heavily relies on the upsampling modules and feature extractors used therein. For the point upsampling module, we propose a novel model called NodeShuffle, which uses a Graph Convolutional Network (GCN) to better encode local point information from point neighborhoods. NodeShuffle is versatile and can be incorporated into any point cloud upsampling pipeline. Extensive experiments show how NodeShuffle consistently improves state-of-the-art upsampling methods. For feature extraction, we also propose a new multi-scale point feature extractor, called Inception DenseGCN. By aggregating features at multiple scales, this feature extractor enables further performance gain in the final upsampled point clouds. We combine Inception DenseGCN with NodeShuffle into a new point upsampling pipeline called PU-GCN. PU-GCN sets new state-of-art performance with much fewer parameters and more efficient inference.
****
Differentiable Patch Selection for Image Recognition
Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, Thomas Unterthiner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2351-2360
Neural Networks require large amounts of memory and compute to process high reso

lution images, even when only a small part of the image is actually informative for the task at hand. We propose a method based on a differentiable Top-K operator to select the most relevant parts of the input to efficiently process high resolution images. Our method may be interfaced with any downstream neural network, is able to aggregate information from different patches in a flexible way, and allows the whole model to be trained end-to-end using backpropagation. We show results for traffic sign recognition, inter-patch relationship reasoning, and fine-grained recognition without using object/part bounding box annotations during training.
****
MaX-DeepLab: End-to-End Panoptic Segmentation With Mask Transformers
Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, Liang-Chieh Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5463-5474
We present MaX-DeepLab, the first end-to-end model for panoptic segmentation. Our approach simplifies the current pipeline that depends heavily on surrogate sub-tasks and hand-designed components, such as box detection, non-maximum suppression, thing-stuff merging, etc. Although these sub-tasks are tackled by area experts, they fail to comprehensively solve the target task. By contrast, our MaX-DeepLab directly predicts class-labeled masks with a mask transformer, and is trained with a panoptic quality inspired loss via bipartite matching. Our mask transformer employs a dual-path architecture that introduces a global memory path in addition to a CNN path, allowing direct communication with any CNN layers. As a result, MaX-DeepLab shows a significant 7.1% PQ gain in the box-free regime on the challenging COCO dataset, closing the gap between box-based and box-free methods for the first time. A small variant of MaX-DeepLab improves 3.0% PQ over DETR with similar parameters and M-Adds. Furthermore, MaX-DeepLab, without test time augmentation, achieves new state-of-the-art 51.3% PQ on COCO test-dev set.
****
Improving Transferability of Adversarial Patches on Face Recognition With Generative Models
Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11845-11854
Face recognition is greatly improved by deep convolutional neural networks (CNNs). Recently, these face recognition models have been used for identity authentication in security sensitive applications. However, deep CNNs are vulnerable to adversarial patches, which are physically realizable and stealthy, raising new security concerns on the real-world applications of these models. In this paper, we evaluate the robustness of face recognition models using adversarial patches based on transferability, where the attacker has limited accessibility to the target models. First, we extend the existing transfer-based attack techniques to generate transferable adversarial patches. However, we observe that the transferability is sensitive to initialization and degrades when the perturbation magnitude is large, indicating the overfitting to the substitute models. Second, we propose to regularize the adversarial patches on the low dimensional data manifold. The manifold is represented by generative models pre-trained on legitimate human face images. Using face-like features as adversarial perturbations through optimization on the manifold, we show that the gaps between the responses of substitute models and the target models dramatically decrease, exhibiting a better transferability. Extensive digital world experiments are conducted to demonstrate the superiority of the proposed method in the black-box setting. We apply the proposed method in the physical world as well.
****
Counterfactual VQA: A Cause-Effect Look at Language Bias
Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, Ji-Rong Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12700-12710
Recent VQA models may tend to rely on language bias as a shortcut and thus fail to sufficiently learn the multi-modal knowledge from both vision and language. I

n this paper, we investigate how to capture and mitigate language bias in VQA. Motivated by causal effects, we proposed a novel counterfactual inference framework, which enables us to capture the language bias as the direct causal effect of questions on answers and reduce the language bias by subtracting the direct language effect from the total causal effect. Experiments demonstrate that our proposed counterfactual inference framework 1) is general to various VQA backbones and fusion strategies, 2) achieves competitive performance on the language-bias sensitive VQA-CP dataset while performs robustly on the balanced VQA v2 dataset without any argumented data. The code is available at https://github.com/yuleiniu/cfvqa.
****
Denoise and Contrast for Category Agnostic Shape Completion
Antonio Alliegro, Diego Valsesia, Giulia Fracastoro, Enrico Magli, Tatiana Tommasi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4629-4638
In this paper, we present a deep learning model that exploits the power of self-supervision to perform 3D point cloud completion, estimating the missing part and a context region around it. Local and global information are encoded in a combined embedding. A denoising pretext task provides the network with the needed local cues, decoupled from the high-level semantics and naturally shared over multiple classes. On the other hand, contrastive learning maximizes the agreement between variants of the same shape with different missing portions, thus producing a representation which captures the global appearance of the shape. The combined embedding inherits category-agnostic properties from the chosen pretext tasks. Differently from existing approaches, this allows to better generalize the completion properties to new categories unseen at training time. Moreover, while decoding the obtained joint representation, we better blend the reconstructed missing part with the partial shape by paying attention to its known surrounding region and reconstructing this frame as auxiliary objective. Our extensive experiments and detailed ablation on the ShapeNet dataset show the effectiveness of each part of the method with new state of the art results. Our quantitative and qualitative analysis confirms how our approach is able to work on novel categories without relying neither on classification and shape symmetry priors, nor on adversarial training procedures.
****
Transformation Invariant Few-Shot Object Detection
Aoxue Li, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3094-3102
Few-shot object detection (FSOD) aims to learn detectors that can be generalized to novel classes with only a few instances. Unlike previous attempts that exploit meta-learning techniques to facilitate FSOD, this work tackles the problem from the perspective of sample expansion. To this end, we propose a simple yet effective Transformation Invariant Principle (TIP) that can be flexibly applied to various meta-learning models for boosting the detection performance on novel class objects. Specifically, by introducing consistency regularization on predictions from various transformed images, we augment vanilla FSOD models with the generalization ability to objects perturbed by various transformation, such as occlusion and noise. Importantly, our approach can extend supervised FSOD models to naturally cope with unlabeled data, thus addressing a more practical and challenging semi-supervised FSOD problem. Extensive experiments on PASCAL VOC and MSCOCO datasets demonstrate the effectiveness of our TIP under both of the two FSOD settings.
****
2D or not 2D? Adaptive 3D Convolution Selection for Efficient Video Recognition
Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, Larry S. Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6155-6164
3D convolutional networks are prevalent for video recognition. While achieving excellent recognition performance on standard benchmarks, they operate on a sequence of frames with 3D convolutions and thus are computationally demanding. Explo

iting large variations among different videos, we introduce Ada3D, a conditional computation framework that learns instance-specific 3D usage policies to determine frames and convolution layers to be used in a 3D network. These policies are derived with a two-head lightweight selection network conditioned on each input video clip. Then, only frames and convolutions that are selected by the selection network are used in the 3D model to generate predictions. The selection network is optimized with policy gradient methods to maximize a reward that encourages making correct predictions with limited computation. We conduct experiments on three video recognition benchmarks and demonstrate that our method achieves similar accuracies to state-of-the-art 3D models while requiring 20%-50% less computation across different datasets. We also show that learned policies are transferable and Ada3D is compatible to different backbones and modern clip selection approaches. Our qualitative analysis indicates that our method allocates fewer 3D convolutions and frames for "static" inputs, yet uses more for motion-intensive clips.
****
Temporal Query Networks for Fine-Grained Video Understanding
Chuhan Zhang, Ankush Gupta, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4486-4496
Our objective in this work is fine-grained classification of actions in untrimmed videos, where the actions may be temporally extended or may span only a few frames of the video. We cast this into a query-response mechanism, where each query addresses a particular question, and has its own response label set. We make the following four contributions: (i) We propose a new model---a Temporal Query Network---which enables the query-response functionality, and a structural understanding of fine-grained actions. It attends to relevant segments for each query with a temporal attention mechanism, and can be trained using only the labels for each query. (ii) We propose a new way---stochastic feature bank update---to train a network on videos of various lengths with the dense sampling required to respond to fine-grained queries. (iii) we compare the TQN to other architectures and text supervision methods, and analyze their pros and cons. Finally, (iv) we evaluate the method extensively on the FineGym and Diving48 benchmarks for fine-grained action classification and surpass the state-of-the-art using only RGB features.
****
Adversarial Generation of Continuous Images
Ivan Skorokhodov, Savva Ignatyev, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10753-10764
In most existing learning systems, images are typically viewed as 2D pixel arrays. However, in another paradigm gaining popularity, a 2D image is represented as an implicit neural representation (INR) -- an MLP that predicts an RGB pixel value given its (x,y) coordinate. In this paper, we propose two novel architectural techniques for building INR-based image decoders: factorized multiplicative modulation and multi-scale INRs, and use them to build a state-of-the-art continuous image GAN. Previous attempts to adapt INRs for image generation were limited to MNIST-like datasets and do not scale to complex real-world data. Our proposed INR-GAN architecture improves the performance of continuous image generators by several times, greatly reducing the gap between continuous image GANs and pixel-based ones. Apart from that, we explore several exciting properties of the INR-based decoders, like out-of-the-box superresolution, meaningful image-space interpolation, accelerated inference of low-resolution images, an ability to extrapolate outside of image boundaries, and strong geometric prior. The project page is located at https://universome.github.io/inr-gan.
****
UniT: Unified Knowledge Transfer for Any-Shot Object Detection and Segmentation
Siddhesh Khandelwal, Raghav Goyal, Leonid Sigal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5951-5961
Methods for object detection and segmentation rely on large scale instance-level annotations for training, which are difficult and time-consuming to collect. Ef

forts to alleviate this look at varying degrees and quality of supervision. Weak
ly-supervised approaches draw on image-level labels to build detectors/segmentor
s, while zero/few-shot methods assume abundant instance-level data for a set of
base classes, and none to a few examples for novel classes. This taxonomy has la
rgely siloed algorithmic designs. In this work, we aim to bridge this divide by
proposing an intuitive and unified semi-supervised model that is applicable to a
 range of supervision: from zero to a few instance-level samples per novel class
. For base classes, our model learns a mapping from weakly-supervised to fully-s
upervised detectors/segmentors. By learning and leveraging visual and lingual si
milarities between the novel and base classes, we transfer those mappings to obt
ain detectors/segmentors for novel classes; refining them with a few novel class
 instance-level annotated samples, if available. The overall model is end-to-end
 trainable and highly flexible. Through extensive experiments on MS-COCO and Pas
cal VOC benchmark datasets we show improved performance in a variety of settings
.
****
Indoor Panorama Planar 3D Reconstruction via Divide and Conquer
Cheng Sun, Chi-Wei Hsiao, Ning-Hsu Wang, Min Sun, Hwann-Tzong Chen; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
21, pp. 11338-11347
Indoor panorama typically consists of human-made structures parallel or perpendi
cular to gravity. We leverage this phenomenon to approximate the scene in a 360-
degree image with (H)orizontal-planes and (V)ertical-planes. To this end, we pro
pose an effective divide-and-conquer strategy that divides pixels based on their
 plane orientation estimation; then, the succeeding instance segmentation module
 conquers the task of planes clustering more easily in each plane orientation gr
oup. Besides, parameters of V-planes depend on camera yaw rotation, but translat
ion-invariant CNNs are less aware of the yaw change. We thus propose a yaw-invar
iant V-planar reparameterization for CNNs to learn. We create a benchmark for in
door panorama planar reconstruction by extending existing 360 depth datasets wit
h ground truth H&V-planes (referred to as "PanoH&V" dataset) and adopt state-of-
the-art planar reconstruction methods to predict H&V-planes as our baselines. Ou
r method outperforms the baselines by a large margin on the proposed dataset.
****
Embedded Discriminative Attention Mechanism for Weakly Supervised Semantic Segme
ntation
Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, Chi Har
old Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2021, pp. 16765-16774
Weakly Supervised Semantic Segmentation (WSSS) with image-level annotation uses
class activation maps from the classifier as pseudo-labels for semantic segmenta
tion. However, such activation maps usually highlight the local discriminative r
egions rather than the whole object, which deviates from the requirement of sema
ntic segmentation. To explore more comprehensive class-specific activation maps,
 we propose an Embedded Discriminative Attention Mechanism (EDAM) by integrating
 the activation map generation into the classification network directly for WSSS
. Specifically, a Discriminative Activation (DA) layer is designed to explicitly
 produce a series of normalized class-specific masks, which are then used to gen
erate class-specific pixel-level pseudo-labels demanded in segmentation. For lea
rning the pseudo-labels, the masks are multiplied with the feature maps after th
e backbone to generate the discriminative activation maps, each of which encodes
 the specific information of the corresponding category in the input images. Giv
en such class-specific activation maps, a Collaborative Multi-Attention (CMA) mo
dule is proposed to extract the collaborative information of each given category
 from images in a batch. In inference, we directly use the activation masks from
 the DA layer as pseudo-labels for segmentation. Based on the generated pseudo-l
abels, we achieve the mIoU of 70:60% on PASCAL VOC 2012 segmentation test set, w
hich is the new state-of-the-art, to our best knowledge. Code and pre-trained mo
dels are available online soon.
****

TextOCR: Towards Large-Scale End-to-End Reasoning for Arbitrary-Shaped Scene Tex
t
Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, Tal Hassner;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2021, pp. 8802-8812
A crucial component for the scene text based reasoning required for TextVQA and
TextCaps datasets involve detecting and recognizing text present in the images u
sing an optical character recognition (OCR) system. The current systems are crip
pled by the unavailability of ground truth text annotations for these datasets a
s well as lack of scene text detection and recognition datasets on real images d
isallowing the progress in the field of OCR and evaluation of scene text based r
easoning in isolation from OCR systems. In this work, we propose TextOCR, an arb
itrary-shaped scene text detection and recognition with 900k annotated words col
lected on real images from TextVQA dataset. We show that current state-of-the-ar
t text-recognition (OCR) models fail to perform well on TextOCR and that trainin
g on TextOCR helps achieve state-of-the-art performance on multiple other OCR da
tasets as well. We use a TextOCR trained OCR model to create PixelM4C model whic
h can do scene text based reasoning on an image in an end-to-end fashion, allowi
ng us to revisit several design choices to achieve new state-of-the-art performa
nce on TextVQA dataset.
****
Distractor-Aware Fast Tracking via Dynamic Convolutions and MOT Philosophy
Zikai Zhang, Bineng Zhong, Shengping Zhang, Zhenjun Tang, Xin Liu, Zhaoxiang Zha
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2021, pp. 1024-1033
A practical long-term tracker typically contains three key properties, i.e., an
efficient model design, an effective global re-detection strategy and a robust d
istractor awareness mechanism. However, most state-of-the-art long-term trackers
 (e.g., Pseudo and re-detecting based ones) do not take all three key properties
 into account and therefore may either be time-consuming or drift to distractors
. To address the issues, we propose a two-task tracking framework (named DMTrack
), which utilizes two core components (i.e., one-shot detection and re-identific
ation (re-id) association) to achieve distractor-aware fast tracking via Dynamic
 convolutions (d-convs) and Multiple object tracking (MOT) philosophy. To achiev
e precise and fast global detection, we construct a lightweight one-shot detecto
r using a novel dynamic convolutions generation method, which provides a unified
 and more flexible way for fusing target information into the search field. To d
istinguish the target from distractors, we resort to the philosophy of MOT to re
ason distractors explicitly by maintaining all potential similarities' tracklets
. Benefited from the strength of high recall detection and explicit object assoc
iation, our tracker achieves state-of-the-art performance on the LaSOT, OxUvA, T
LP, VOT2018LT and VOT2019LT benchmarks and runs in real-time (3x faster than com
parisons).
****
Scaling Local Self-Attention for Parameter Efficient Visual Backbones
Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtm
an, Jonathon Shlens; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2021, pp. 12894-12904
Self-attention has the promise of improving computer vision systems due to param
eter-independent scaling of receptive fields and content-dependent interactions,
 in contrast to parameter-dependent scaling and content-independent interactions
 of convolutions. Self-attention models have recently been shown to have encoura
ging improvements on accuracy-parameter trade-offs compared to baseline convolut
ional models such as ResNet-50. In this work, we develop self-attention models t
hat can outperform not just the canonical baseline models, but even the high-per
forming convolutional models. We propose two extensions to self-attention that,
in conjunction with a more efficient implementation of self-attention, improve t
he speed, memory usage, and accuracy of these models. We leverage these improvem
ents to develop a new self-attention model family, HaloNets, which reach state-o
f-the-art accuracies on the parameter-limited setting of the ImageNet classifica

tion benchmark. In preliminary transfer learning experiments, we find that HaloNet models outperform much larger models and have better inference performance. On harder tasks such as object detection and instance segmentation, our simple local self-attention and convolutional hybrids show improvements over very strong baselines. These results mark another step in demonstrating the efficacy of self-attention models on settings traditionally dominated by convolutions.
****

Image Inpainting Guided by Coherence Priors of Semantics and Textures
Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, Shin'ichi Satoh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6539-6548
Existing inpainting methods have achieved promising performance in recovering defected images of specific scenes. However, filling holes involving multiple semantic categories remains challenging due to the obscure semantic boundaries and the mixture of different semantic textures. In this paper, we introduce coherence priors between the semantics and textures which make it possible to concentrate on completing separate textures in a semantic-wise manner. Specifically, we adopt a multi-scale joint optimization framework to first model the coherence priors and then accordingly interleavingly optimize image inpainting and semantic segmentation in a coarse-to-fine manner. A Semantic-Wise Attention Propagation (SWAP) module is devised to refine completed image textures across scales by exploring non-local semantic coherence, which effectively mitigates mix-up of textures. We also propose two coherence losses to constrain the consistency between the semantics and the inpainted image in terms of the overall structure and detailed textures. Experimental results demonstrate the superiority of our proposed method for challenging cases with complex holes.
****

Multi-Source Domain Adaptation With Collaborative Learning for Semantic Segmentation
Jianzhong He, Xu Jia, Shuaijun Chen, Jianzhuang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11008-11017
Multi-source unsupervised domain adaptation (MSDA) aims at adapting models trained on multiple labeled source domains to an unlabeled target domain. In this paper, we propose a novel multi-source domain adaptation framework based on collaborative learning for semantic segmentation. Firstly, a simple image translation method is introduced to align the pixel value distribution to reduce the gap between source domains and target domain to some extent. Then, to fully exploit the essential semantic information across source domains, we propose a collaborative learning method for domain adaptation without seeing any data from target domain. In addition, similar to the setting of unsupervised domain adaptation, unlabeled target domain data is leveraged to further improve the performance of domain adaptation. This is achieved by additionally constraining the outputs of multiple adaptation models with pseudo labels online generated by an ensembled model. Extensive experiments and ablation studies are conducted on the widely-used domain adaptation benchmark datasets in semantic segmentation. Our proposed method achieves 59.0% mIoU on the validation set of Cityscapes by training on the labeled Synscapes and GTA5 datasets and unlabeled training set of Cityscapes. It significantly outperforms all previous state-of-the-arts single-source and multi-source unsupervised domain adaptation methods.
****

Positive-Congruent Training: Towards Regression-Free Model Updates
Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14299-14308
Reducing inconsistencies in the behavior of different versions of an AI system can be as important in practice as reducing its overall error. In image classification, sample-wise inconsistencies appear as "negative flips": A new model incorrectly predicts the output for a test sample that was correctly classified by the old (reference) model. Positive-congruent (PC) training aims at reducing error

rate while at the same reducing negative flips, thus maximizing congruency with the reference model only on positive predictions, unlike model distillation. We propose a simple approach for PC training, Focal Distillation, which enforces congruence with the reference model by giving more weights to samples that were correctly classified. We also found that, if the reference model itself can be chosen as an ensemble of multiple deep neural networks, negative flips can be further reduced without affecting the new model's accuracy.
****
FrameExit: Conditional Early Exiting for Efficient Video Recognition
Amir Ghodrati, Babak Ehteshami Bejnordi, Amirhossein Habibian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15608-15618
In this paper, we propose a conditional early exiting framework for efficient video recognition. While existing works focus on selecting a subset of salient frames to reduce the computation costs, we propose to use a simple sampling strategy combined with conditional early exiting to enable efficient recognition. Our model automatically learns to process fewer frames for simpler videos and more frames for complex ones. To achieve this, we employ a cascade of gating modules to automatically determine the earliest point in processing where an inference is sufficiently reliable. We generate on-the-fly supervision signals to the gates to provide a dynamic trade-off between accuracy and computational cost. Our proposed model outperforms competing methods on three large-scale video benchmarks. In particular, on ActivityNet1.3 and mini-kinetics, we outperform the state-of-the-art efficient video recognition methods with 1.3x and 2.1x less GFLOPs, respectively. Additionally, our method sets a new state of the art for efficient video understanding on the HVU benchmark.
****
Neighbor2Neighbor: Self-Supervised Denoising From Single Noisy Images
Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, Jianzhuang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14781-14790
In the last few years, image denoising has benefited a lot from the fast development of neural networks. However, the requirement of large amounts of noisy-clean image pairs for supervision limits the wide use of these models. Although there have been a few attempts in training an image denoising model with only single noisy images, existing self-supervised denoising approaches suffer from inefficient network training, loss of useful information, or dependence on noise modeling. In this paper, we present a very simple yet effective method named Neighbor2Neighbor to train an effective image denoising model with only noisy images. Firstly, a random neighbor sub-sampler is proposed for the generation of training image pairs. In detail, input and target used to train a network are images sub-sampled from the same noisy image, satisfying the requirement that paired pixels of paired images are neighbors and have very similar appearance with each other. Secondly, a denoising network is trained on sub-sampled training pairs generated in the first stage, with a proposed regularizer as additional loss for better performance. The proposed Neighbor2Neighbor framework is able to enjoy the progress of state-of-the-art supervised denoising networks in network architecture design. Moreover, it avoids heavy dependence on the assumption of the noise distribution. We explain our approach from a theoretical perspective and further validate it through extensive experiments, including synthetic experiments with different noise distributions in sRGB space and real-world experiments on a denoising benchmark dataset in raw-RGB space.
****
Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing
Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1622-1631
To address the challenging task of instance-aware human part parsing, a new bottom-up regime is proposed to learn category-level human semantic segmentation as

well as multi-person pose estimation in a joint and end-to-end manner. It is a c
ompact, efficient and powerful framework that exploits structural information ov
er different human granularities and eases the difficulty of person partitioning
. Specifically, a dense-to-sparse projection field, which allows explicitly asso
ciating dense human semantics with sparse keypoints, is learnt and progressively
 improved over the network feature pyramid for robustness. Then, the difficult p
ixel grouping problem is cast as an easier, multi-person joint assembling task.
By formulating joint association as maximum-weight bipartite matching, a differe
ntiable solution is developed to exploit projected gradient descent and Dykstra'
s cyclic projection algorithm. This makes our method end-to-end trainable and al
lows back-propagating the grouping error to directly supervise multi-granularity
 human representation learning. This is distinguished from current bottom-up hum
an parsers or pose estimators which require sophisticated post-processing or heu
ristic greedy algorithms. Experiments on three instance-aware human parsing data
sets show that our model outperforms other bottom-up alternatives with much more
 efficient inference.
****
Dynamic Weighted Learning for Unsupervised Domain Adaptation
Ni Xiao, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2021, pp. 15242-15251
Unsupervised domain adaptation (UDA) aims to improve the classification performa
nce on an unlabeled target domain by leveraging information from a fully labeled
 source domain. Recent approaches explore domain-invariant and class-discriminan
t representations to tackle this task. These methods, however, ignore the intera
ction between domain alignment learning and class discrimination learning. As a
result, the missing or inadequate tradeoff between domain alignment and class di
scrimination are prone to the problem of negative transfer. In this paper, we pr
opose Dynamic Weighted Learning (DWL) to avoid the discriminability vanishing pr
oblem caused by excessive alignment learning and domain misalignment problem cau
sed by excessive discriminant learning. Technically, DWL dynamically weights the
 learning losses of alignment and discriminability by introducing the degree of
alignment and discriminability. Besides, the problem of sample imbalance across
domains is first considered in our work, and we solve the problem by weighing th
e samples to guarantee information balance across domains. Extensive experiments
 demonstrate that DWL has an excellent performance in several benchmark datasets
.
****
Using Shape To Categorize: Low-Shot Learning With an Explicit Shape Bias
Stefan Stojanov, Anh Thai, James M. Rehg; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1798-1808
It is widely accepted that reasoning about object shape is important for object
recognition. However, the most powerful object recognition methods today do not
explicitly make use of object shape during learning. In this work, motivated by
recent developments in low-shot learning, findings in developmental psychology,
and the increased use of synthetic data in computer vision research, we investig
ate how reasoning about 3D shape can be used to improve low-shot learning method
s' generalization performance. We propose a new way to improve existing low-shot
 learning approaches by learning a discriminative embedding space using 3D objec
t shape, and using this embedding by learning how to map images into it. Our new
 approach improves the performance of image-only low-shot learning approaches on
 multiple datasets. We also introduce Toys4K, a 3D object dataset with the large
st number of object categories currently available, which supports low-shot lear
ning.
****
Face Forensics in the Wild
Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, Jianbing Shen; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 57
78-5788
On existing public benchmarks, face forgery detection techniques have achieved g
reat success. However, when used in multi-person videos, which often contain man

y people active in the scene with only a small subset having been manipulated, their performance remains far from being satisfactory. To take face forgery detection to a new level, we construct a novel large-scale dataset, called FFIW-10K, which comprises 10,000 high-quality forgery videos, with an average of three human faces in each frame. The manipulation procedure is fully automatic, controlled by a domain-adversarial quality assessment network, making our dataset highly scalable with low human cost. In addition, we propose a novel algorithm to tackle the task of multi-person face forgery detection. Supervised by only video-level label, the algorithm explores multiple instance learning and learns to automatically attend to tampered faces. Our algorithm outperforms representative approaches for both forgery classification and localization on FFIW-10K, and also shows high generalization ability on existing benchmarks. We hope that our dataset and study will help the community to explore this new field in more depth.
****
Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain
Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 772-781
The remarkable success in face forgery techniques has received considerable attention in computer vision due to security concerns. We observe that up-sampling is a necessary step of most face forgery techniques, and cumulative up-sampling will result in obvious changes in the frequency domain, especially in the phase spectrum. According to the property of natural images, the phase spectrum preserves abundant frequency components that provide extra information and complement the loss of the amplitude spectrum. To this end, we present a novel Spatial-Phase Shallow Learning (SPSL) method, which combines spatial image and phase spectrum to capture the up-sampling artifacts of face forgery to improve the transferability, for face forgery detection. And we also theoretically analyze the validity of utilizing the phase spectrum. Moreover, we notice that local texture information is more crucial than high-level semantic information for the face forgery detection task. So we reduce the receptive fields by shallowing the network to suppress high-level features and focus on the local region. Extensive experiments show that SPSL can achieve the state-of-the-art performance on cross-datasets evaluation as well as multi-class classification and obtain comparable results on single dataset evaluation.
****
A Closer Look at Fourier Spectrum Discrepancies for CNN-Generated Images Detection
Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7200-7209
CNN-based generative modelling has evolved to produce synthetic images indistinguishable from real images in the RGB pixel space. Recent works have observed that CNN-generated images share a systematic shortcoming in replicating high frequency Fourier spectrum decay attributes. Furthermore, these works have successfully exploited this systematic shortcoming to detect CNN-generated images reporting up to 99% accuracy across multiple state-of-the-art GAN models. In this work, we investigate the validity of assertions claiming that CNN-generated images are unable to achieve high frequency spectral decay consistency. We meticulously construct a counterexample space of high frequency spectral decay consistent CNN-generated images emerging from our handcrafted experiments using DCGAN, LSGAN, WGAN-GP and StarGAN, where we empirically show that this frequency discrepancy can be avoided by a minor architecture change in the last upsampling operation. We subsequently use images from this counterexample space to successfully bypass the recently proposed forensics detector which leverages on high frequency Fourier spectrum decay attributes for CNN-generated image detection. Through this study, we show that high frequency Fourier spectrum decay discrepancies are not inherent characteristics for existing CNN-based generative models---contrary to the belief of some existing work---, and such features are not robust to perform synth

etic image detection. Our results prompt re-thinking of using high frequency Fourier spectrum decay attributes for CNN-generated image detection. Code and models are available at https://keshik6.github.io/Fourier-Discrepancies-CNN-Detection/
****
Learning Delaunay Surface Elements for Mesh Reconstruction
Marie-Julie Rakotosaona, Paul Guerrero, Noam Aigerman, Niloy J. Mitra, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 22-31
We present a method for reconstructing triangle meshes from point clouds. Existing learning-based methods for mesh reconstruction mostly generate triangles individually, making it hard to create manifold meshes. We leverage the properties of 2D Delaunay triangulations to construct a mesh from manifold surface elements. Our method first estimates local geodesic neighborhoods around each point. We then perform a 2D projection of these neighborhoods using a learned logarithmic map. A Delaunay triangulation in this 2D domain is guaranteed to produce a manifold patch, which we call a surface element. We synchronize the local 2D projections of neighboring elements to maximize the manifoldness of the reconstructed mesh. Our results show that we achieve better overall manifoldness of our reconstructed meshes than current methods to reconstruct meshes with arbitrary topology. Our code, data and pretrained models can be found online: https://github.com/mrakotosaon/dse-meshing
****
FaceSec: A Fine-Grained Robustness Evaluation Framework for Face Recognition Systems
Liang Tong, Zhengzhang Chen, Jingchao Ni, Wei Cheng, Dongjin Song, Haifeng Chen, Yevgeniy Vorobeychik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13254-13263
We present FACESEC, a framework for fine-grained robustness evaluation of face recognition systems. FACESEC evaluation is performed along four dimensions of adversarial modeling: the nature of perturbation (e.g., pixel-level or face accessories), the attacker's system knowledge (about training data and learning architecture), goals (dodging or impersonation), and capability (tailored to individual inputs or across sets of these). We use FACESEC to study five face recognition systems in both closed-set and open-set settings, and to evaluate the state-of-the-art approach for defending against physically realizable attacks on these. We find that accurate knowledge of neural architecture is significantly more important than knowledge of the training data in black-box attacks. Moreover, we observe that open-set face recognition systems are more vulnerable than closed-set systems under different types of attacks. The efficacy of attacks for other threat model variations, however, appears highly dependent on both the nature of perturbation and the neural network architecture. For example, attacks that involve adversarial face masks are usually more potent, even against adversarially trained models, and the ArcFace architecture tends to be more robust than the others.
****
Dynamic Head: Unifying Object Detection Heads With Attentions
Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7373-7382
The complex nature of combining localization and classification in object detection has resulted in the flourished development of methods. Previous works tried to improve the performance in various object detection heads but failed to present a unified view. In this paper, we present a novel dynamic head framework to unify object detection heads with attentions. By coherently combining multiple self-attention mechanisms between feature levels for scale-awareness, among spatial locations for spatial-awareness, and within output channels for task-awareness, the proposed approach significantly improves the representation ability of object detection heads without any computational overhead. Further experiments demonstrate that the effectiveness and efficiency of the proposed dynamic head on the COCO benchmark. With a standard ResNeXt-101-DCN backbone, we largely improve t

he performance over popular object detectors and achieve a new state-of-the-art at 54.0 AP. The code will be released at https://github.com/microsoft/DynamicHead.
****
Riggable 3D Face Reconstruction via In-Network Optimization
Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6216-6225
This paper presents a method for riggable 3D face reconstruction from monocular images, which jointly estimates a personalized face rig and per-image parameters including expressions, poses, and illuminations. To achieve this goal, we design an end-to-end trainable network embedded with a differentiable in-network optimization. The network first parameterizes the face rig as a compact latent code with a neural decoder, and then estimates the latent code as well as per-image parameters via a learnable optimization. By estimating a personalized face rig, our method goes beyond static reconstructions and enables downstream applications such as video retargeting. In-network optimization explicitly enforces constraints derived from the first principles, thus introduces additional priors than regression-based methods. Finally, data-driven priors from deep learning are utilized to constrain the ill-posed monocular setting and ease the optimization difficulty. Experiments demonstrate that our method achieves SOTA reconstruction accuracy, reasonable robustness and generalization ability, and supports standard face rig applications.
****
One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing
Ting-Chun Wang, Arun Mallya, Ming-Yu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10039-10049
We propose a neural talking-head video synthesis model and demonstrate its application to video conferencing. Our model learns to synthesize a talking-head video using a source image containing the target person's appearance and a driving video that dictates the motion in the output. Our motion is encoded based on a novel keypoint representation, where the identity-specific and motion-related information is decomposed unsupervisedly. Extensive experimental validation shows that our model outperforms competing methods on benchmark datasets. Moreover, our compact keypoint representation enables a video conferencing system that achieves the same visual quality as the commercial H.264 standard while only using one-tenth of the bandwidth. Besides, we show our keypoint representation allows the user to rotate the head during synthesis, which is useful for simulating face-to-face video conferencing experiences.
****
S2R-DepthNet: Learning a Generalizable Depth-Specific Structural Representation
Xiaotian Chen, Yuwang Wang, Xuejin Chen, Wenjun Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3034-3043
Human can infer the 3D geometry of a scene from a sketch instead of a realistic image, which indicates that the spatial structure plays a fundamental role in understanding the depth of scenes. We are the first to explore the learning of a depth-specific structural representation, which captures the essential feature for depth estimation and ignores irrelevant style information. Our S2R-DepthNet (Synthetic to Real DepthNet) can be well generalized to unseen real-world data directly even though it is only trained on synthetic data. S2R-DepthNet consists of: a) a Structure Extraction (STE) module which extracts a domaininvariant structural representation from an image by disentangling the image into domain-invariant structure and domain-specific style components, b) a Depth-specific Attention (DSA) module, which learns task-specific knowledge to suppress depth-irrelevant structures for better depth estimation and generalization, and c) a depth prediction module (DP) to predict depth from the depth-specific representation. Without access of any real-world images, our method even outperforms the state-of-the-art unsupervised domain adaptation methods which use real-world images of the target domain for training. In addition, when using a small amount of labeled real-world data, we achieve the state-of-the-art performance under the semi-supervi

sed setting.
****
Holistic 3D Human and Scene Mesh Estimation From Single View Images
Zhenzhen Weng, Serena Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 334-343
The 3D world limits the human body pose and the human body pose conveys information about the surrounding objects. Indeed, from a single image of a person placed in an indoor scene, we as humans are adept at resolving ambiguities of the human pose and room layout through our knowledge of the physical laws and prior perception of the plausible object and human poses. However, few computer vision models fully leverage this fact. In this work, we propose a holistically trainable model that perceives the 3D scene from a single RGB image, estimates the camera pose and the room layout, and reconstructs both human body and object meshes. By imposing a set of comprehensive and sophisticated losses on all aspects of the estimations, we show that our model outperforms existing human body mesh methods and indoor scene reconstruction methods. To the best of our knowledge, this is the first model that outputs both object and human predictions at the mesh level, and performs joint optimization on the scene and human poses.
****
MIST: Multiple Instance Spatial Transformer
Baptiste Angles, Yuhe Jin, Simon Kornblith, Andrea Tagliasacchi, Kwang Moo Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2412-2422
We propose a deep network that can be trained to tackle image reconstruction and classification problems that involve detection of multiple object instances, without any supervision regarding their whereabouts. The network learns to extract the most significant top-K patches, and feeds these patches to a task-specific network -- e.g., auto-encoder or classifier -- to solve a domain specific problem. The challenge in training such a network is the non-differentiable top-K selection process. To address this issue, we lift the training optimization problem by treating the result of top-K selection as a slack variable, resulting in a simple, yet effective, multi-stage training. Our method is able to learn to detect recurrent structures in the training dataset by learning to reconstruct images. It can also learn to localize structures when only knowledge on the occurrence of the object is provided, and in doing so it outperforms the state-of-the-art.
****
FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation
Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3003-3013
In this work, we present FFB6D, a full flow bidirectional fusion network designed for 6D pose estimation from a single RGBD image. Our key insight is that appearance information in the RGB image and geometry information from the depth image are two complementary data sources, and it still remains unknown how to fully leverage them. Towards this end, we propose FFB6D, which learns to combine appearance and geometry information for representation learning as well as output representation selection. Specifically, at the representation learning stage, we build bidirectional fusion modules in the full flow of the two networks, where fusion is applied to each encoding and decoding layer. In this way, the two networks can leverage local and global complementary information from the other one to obtain better representations. Moreover, at the output representation stage, we designed a simple but effective 3D keypoints selection algorithm considering the texture and geometry information of objects, which simplifies keypoint localization for precise pose estimation. Experimental results show that our method outperforms the state-of-the-art by large margins on several benchmarks. The code of this work will be open-source to the community.
****
Shape From Sky: Polarimetric Normal Recovery Under the Sky
Tomoki Ichikawa, Matthew Purri, Ryo Kawahara, Shohei Nobuhara, Kristin Dana, Ko Nishino; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R

ecognition (CVPR), 2021, pp. 14832-14841
The sky exhibits a unique spatial polarization pattern by scattering the unpolarized sun light. Just like insects use this unique angular pattern to navigate, we use it to map pixels to directions on the sky. That is, we show that the unique polarization pattern encoded in the polarimetric appearance of an object captured under the sky can be decoded to reveal the surface normal at each pixel. We derive a polarimetric reflection model of a diffuse plus mirror surface lit by the sun and a clear sky. This model is used to recover the per-pixel surface normal of an object from a single polarimetric image or from multiple polarimetric images captured under the sky at different times of the day. We experimentally evaluate the accuracy of our shape-from-sky method on a number of real objects of different surface compositions. The results clearly show that this passive approach to fine-geometry recovery that fully leverages the unique illumination made by nature is a viable option for 3D sensing. With the advent of quad-Bayer polarization chips, we believe the implications of our method span a wide range of domains.
****
Adversarially Adaptive Normalization for Single Domain Generalization
Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, Mingyuan Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8208-8217
Single domain generalization aims to learn a model that performs well on many unseen domains with only one domain data for training. Existing works focus on studying the adversarial domain augmentation (ADA) to improve the model's generalization capability. The impact on domain generalization from the statistics of normalization layers is still underinvestigated. In this paper, we propose a generic normalization approach, adaptive standardization and rescaling normalization (ASR-Norm), to complement the missing part in the previous works. ASR-Norm learns both the standardization and rescaling statistics via neural networks. This new form of normalization can be viewed as a generic form of traditional normalizations. When trained with ADA, the statistics in ASR-Norm are learned to be adaptive to the data coming from different domains, and hence improves the model generalization performance across domains, especially on the target domain with large discrepancy from the source domain. The experimental results show that ASR-Norm can bring consistent improvement to the state-of-the-art ADA approaches by 1.6%, 2.7%, and 6.3% averagely on the Digits, CIFAR-10-C, and PACS benchmarks, respectively. As a generic tool, the improvement introduced by ASR-Norm is agnostic to the choice of ADA methods.
****
Rethinking Channel Dimensions for Efficient Model Design
Dongyoon Han, Sangdoo Yun, Byeongho Heo, YoungJoon Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 732-741
Designing an efficient model within the limited computational cost is challenging. We argue the accuracy of a lightweight model has been further limited by the design convention: a stage-wise configuration of the channel dimensions, which looks like a piecewise linear function of the network stage. In this paper, we study an effective channel dimension configuration towards better performance than the convention. To this end, we empirically study how to design a single layer properly by analyzing the rank of the output feature. We then investigate the channel configuration of a model by searching network architectures concerning the channel configuration under the computational cost restriction. Based on the investigation, we propose a simple yet effective channel configuration that can be parameterized by the layer index. As a result, our proposed model following the channel parameterization achieves remarkable performance on ImageNet classification and transfer learning tasks including COCO object detection, COCO instance segmentation, and fine-grained classifications. Code and ImageNet pretrained models are available at https://github.com/clovaai/rexnet.
****
A Self-Boosting Framework for Automated Radiographic Report Generation

Zhanyu Wang, Luping Zhou, Lei Wang, Xiu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2433-2442
Automated radiographic report generation is a challenging task since it requires to generate paragraphs describing fine-grained visual differences of cases, especially for those between the diseased and the healthy. Existing image captioning methods commonly target at generic images, and lack mechanism to meet this requirement. To bridge this gap, in this paper, we propose a self-boosting framework that improves radiographic report generation based on the cooperation of the main task of report generation and anauxiliary task of image-text matching. The two tasks are built as the two branches of a network model and influence each other in a cooperative way. On one hand, the image-text matching branch helps to learn highly text-correlated visual features for the report generation branch to output high quality reports. One the other hand, the improved reports produced by the report generation branch provideadditional harder samples for the image-text matching task and enforce the latter to improve itself by learning better visual and text feature representations. This, in turn, helps improve the report generation branch again. These two branches are jointly trained to help improve each other iteratively and progressively, so that the whole model is self-boosted without requiring any external resources. Additionally, in the loss function, our model evaluates the quality of the generated reports not only on the word similarity as common approaches do (via minimizing a cross-entropy loss), but also on the feature similarity at high-level, while the latter is provided by the text-encoder of the image-text matching branch. Experimental results demonstrate the effectiveness of our method on two public datasets, showing its superior performance over other state-of-the-art medical report generation methods.
****
RAFT-3D: Scene Flow Using Rigid-Motion Embeddings
Zachary Teed, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8375-8384
We address the problem of scene flow: given a pair of stereo or RGB-D video frames, estimate pixelwise 3D motion. We introduce RAFT-3D, a new deep architecture for scene flow. RAFT-3D is based on the RAFT model developed for optical flow but iteratively updates a dense field of pixelwise SE3 motion instead of 2D motion. A key innovation of RAFT-3D is rigid-motion embeddings, which represent a soft grouping of pixels into rigid objects. Integral to rigid-motion embeddings is Dense-SE3, a differentiable layer that enforces geometric consistency of the embeddings. Experiments show that RAFT-3D achieves state-of-the-art performance. On FlyingThings3D, under the two-view evaluation, we improved the best published accuracy (delta < 0.05) from 34.3% to 83.7%. On KITTI, we achieve an error of 5.77, outperforming the best published method (6.31), despite using no object instance supervision.
****
Orthogonal Over-Parameterized Training
Weiyang Liu, Rongmei Lin, Zhen Liu, James M. Rehg, Liam Paull, Li Xiong, Le Song, Adrian Weller; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7251-7260
The inductive bias of a neural network is largely determined by the architecture and the training algorithm. To achieve good generalization, how to effectively train a neural network is of great importance. We propose a novel orthogonal over-parameterized training (OPT) framework that can provably minimize the hyperspherical energy which characterizes the diversity of neurons on a hypersphere. By maintaining the minimum hyperspherical energy during training, OPT can greatly improve the empirical generalization. Specifically, OPT fixes the randomly initialized weights of the neurons and learns an orthogonal transformation that applies to these neurons. We consider multiple ways to learn such an orthogonal transformation, including unrolling orthogonalization algorithms, applying orthogonal parameterization, and designing orthogonality-preserving gradient descent. For better scalability, we propose the stochastic OPT which performs orthogonal transformation stochastically for partial dimensions of neurons. Interestingly, OPT reveals that learning a proper coordinate system for neurons is crucial to genera

lization. We provide some insights on why OPT yields better generalization. Extensive experiments validate the superiority of OPT over the standard training.
****
Masksembles for Uncertainty Estimation
Nikita Durasov, Timur Bagautdinov, Pierre Baque, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13539-13548
Deep neural networks have amply demonstrated their prowess but estimating the reliability of their predictions remains challenging. Deep Ensembles are widely considered as being one of the best methods for generating uncertainty estimates but are very expensive to train and evaluate. MC-Dropout is another popular alternative, which is less expensive, but also less reliable. Our central intuition is that there is a continuous spectrum of ensemble-like models of which MC-Dropout and Deep Ensembles are extreme examples. The first uses effectively infinite number of highly correlated models while the second relies on a finite number of independent models. To combine the benefits of both, we introduce Masksembles. Instead of randomly dropping parts of the network as in MC-dropout, Masksemble relies on a fixed number of binary masks, which are parameterized in a way that allows to change correlations between individual models. Namely, by controlling the overlap between the masks and their density one can choose the optimal configuration for the task at hand. This leads to a simple and easy to implement method with performance on par with Ensembles at a fraction of the cost. We experimentally validate Masksembles on two widely used datasets, CIFAR10 and ImageNet.
****
Network Pruning via Performance Maximization
Shangqian Gao, Feihu Huang, Weidong Cai, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9270-9280
Channel pruning is a class of powerful methods for model compression. When pruning a neural network, it's ideal to obtain a sub-network with higher accuracy. However, a sub-network does not necessarily have high accuracy with low classification loss (loss-metric mismatch). In the paper, we first consider the loss-metric mismatch problem for pruning and propose a novel channel pruning method for Convolutional Neural Networks (CNNs) by directly maximizing the performance (i.e., accuracy) of sub-networks. Specifically, we train a stand-alone neural network to predict sub-networks' performance and then maximize the output of the network as a proxy of accuracy to guide pruning. Training such a performance prediction network efficiently is not an easy task, and it may potentially suffer from the problem of catastrophic forgetting and the imbalance distribution of sub-networks. To deal with this challenge, we introduce a corresponding episodic memory to update and collect sub-networks during the pruning process. In the experiment section, we further demonstrate that the gradients from the performance prediction network and the classification loss have different directions. Extensive experimental results show that the proposed method can achieve state-of-the-art performance with ResNet, MobileNetV2, and ShuffleNetV2+ on ImageNet and CIFAR-10.
****
Closing the Loop: Joint Rain Generation and Removal via Disentangled Image Translation
Yuntong Ye, Yi Chang, Hanyu Zhou, Luxin Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2053-2062
Existing deep learning-based image deraining methods have achieved promising performance for synthetic rainy images, typically rely on the pairs of sharp images and simulated rainy counterparts. However, these methods suffer from significant performance drop when facing the real rain, because of the huge gap between the simplified synthetic rain and the complex real rain. In this work, we argue that the rain generation and removal are the two sides of the same coin and should be tightly coupled. To close the loop, we propose to jointly learn real rain generation and removal procedure within a unified disentangled image translation framework. Specifically, we propose a bidirectional disentangled translation network, in which each unidirectional network contains two loops of joint rain gener

ation and removal for both the real and synthetic rain image, respectively. Mean while, we enforce the disentanglement strategy by decomposing the rainy image in to a clean background and rain layer (rain removal), in order to better preserve the identity background via both the cycle-consistency loss and adversarial los s, and ease the rain layer translating between the real and synthetic rainy imag e. A counterpart composition with the entanglement strategy is symmetrically app lied for rain generation. Extensive experiments on synthetic and real-world rain datasets show the superiority of proposed method compared to state-of-the-arts.
****
ACTION-Net: Multipath Excitation for Action Recognition
Zhengwei Wang, Qi She, Aljosa Smolic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13214-13223
Spatial-temporal, channel-wise, and motion patterns are three complementary and crucial types of information for video action recognition. Conventional 2D CNNs are computationally cheap but cannot catch temporal relationships; 3D CNNs can a chieve good performance but are computationally intensive. In this work, we tack le this dilemma by designing a generic and effective module that can be embedded into 2D CNNs. To this end, we propose a spAtio-temporal, Channel and moTion exc itatION (ACTION) module consisting of three paths: Spatio-Temporal Excitation (S TE) path, Channel Excitation (CE) path, and Motion Excitation (ME) path. The STE path employs one channel 3D convolution to characterize spatio-temporal represe ntation. The CE path adaptively recalibrates channel-wise feature responses by e xplicitly modeling interdependencies between channels in terms of the temporal a spect. The ME path calculates feature-level temporal differences, which is then utilized to excite motion-sensitive channels. We equip 2D CNNs with the proposed ACTION module to form a simple yet effective ACTION-Net with very limited extra computational cost. ACTION-Net is demonstrated by consistently outperforming 2D CNN counterparts on three backbones (i.e., ResNet-50, MobileNet V2 and BNIncept ion) employing three datasets (i.e., Something-Something V2, Jester, and EgoGest ure). Code is provided at https://github.com/V-Sense/ACTION-Net.
****
Co-Attention for Conditioned Image Matching
Olivia Wiles, Sebastien Ehrhardt, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15920-15 929
We propose a new approach to determine correspondences between image pairs in th e wild under large changes in illumination, viewpoint, context, and material. Wh ile other approaches find correspondences between pairs of images by treating th e images independently, we instead condition on both images to implicitly take a ccount of the differences between them. To achieve this, we introduce (i) a spat ial attention mechanism (a co-attention module, CoAM) for conditioning the learn ed features on both images, and (ii) a distinctiveness score used to choose the best matches at test time. CoAM can be added to standard architectures and train ed using self-supervision or supervised data, and achieves a significant perform ance improvement under hard conditions, e.g. large viewpoint changes. We demonst rate that models using CoAM achieve state-of-the-art or competitive results on a wide range of tasks: local matching, camera localization, 3D reconstruction, an d image stylization.
****
EventZoom: Learning To Denoise and Super Resolve Neuromorphic Events
Peiqi Duan, Zihao W. Wang, Xinyu Zhou, Yi Ma, Boxin Shi; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 128 24-12833
We address the problem of jointly denoising and super resolving neuromorphic eve nts, a novel visual signal that represents thresholded temporal gradients in a s pace-time window. The challenge for event signal processing is that they are asy nchronously generated, and do not carry absolute intensity but only binary signs informing temporal variations. To study event signal formation and degradation, we implement a display-camera system which enables multi-resolution event recor ding. We further propose EventZoom, a deep neural framework with a backbone arch

itecture of 3D U-Net. EventZoom is trained in a noise-to-noise fashion where the two ends of the network are unfiltered noisy events, enforcing noise-free event restoration. For resolution enhancement, EventZoom incorporates an event-to-image module supervised by high resolution images. Our results showed that EventZoom achieves at least 40x temporal efficiency compared to state-of-the-art event denoisers. Additionally, we demonstrate that EventZoom enables performance improvements on applications including event-based visual object tracking and image reconstruction. EventZoom achieves state-of-the-art super resolved image reconstruction results while being 10x faster.
****

Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, Sanghyuk Chun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2340-2350

ImageNet has been the most popular image classification benchmark, but it is also the one with a significant level of label noise. Recent studies have shown that many samples contain multiple classes, despite being assumed to be a single-label benchmark. They have thus proposed to turn ImageNet evaluation into a multi-label task, with exhaustive multi-label annotations per image. However, they have not fixed the training set, presumably because of a formidable annotation cost. We argue that the mismatch between single-label annotations and effectively multi-label images is equally, if not more, problematic in the training setup, where random crops are applied. With the single-label annotations, a random crop of an image may contain an entirely different object from the ground truth, introducing noisy or even incorrect supervision during training. We thus re-label the ImageNet training set with multi-labels. We address the annotation cost barrier by letting a strong image classifier, trained on an extra source of data, generate the multi-labels. We utilize the pixel-wise multi-label predictions before the final pooling layer, in order to exploit the additional location-specific supervision signals. Training on the re-labeled samples results in improved model performances across the board. ResNet-50 attains the top-1 accuracy of 78.9% on ImageNet with our localized multi-labels, which can be further boosted to 80.2% with the CutMix regularization. We show that the models trained with localized multi-labels also outperform the baselines on transfer learning to object detection and instance segmentation tasks, and various robustness benchmarks. The re-labeled ImageNet training set, pre-trained weights, and the source code are available at https://github.com/naver-ai/relabel_imagenet.
****

CoCosNet v2: Full-Resolution Correspondence Learning for Image Translation

Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11465-11475

We present the full-resolution correspondence learning for cross-domain images, which aids image translation. We adopt a hierarchical strategy that uses the correspondence from coarse level to guide the fine levels. At each hierarchy, the correspondence can be efficiently computed via PatchMatch that iteratively leverages the matchings from the neighborhood. Within each PatchMatch iteration, the ConvGRU module is employed to refine the current correspondence considering not only the matchings of larger context but also the historic estimates. The proposed CoCosNet v2, a GRU-assisted PatchMatch approach, is fully differentiable and highly efficient. When jointly trained with image translation, full-resolution semantic correspondence can be established in an unsupervised manner, which in turn facilitates the exemplar-based image translation. Experiments on diverse translation tasks show that CoCosNet v2 performs considerably better than state-of-the-art literature on producing high-resolution images.
****

SceneGraphFusion: Incremental 3D Scene Graph Prediction From RGB-D Sequences

Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (

CVPR), 2021, pp. 7515-7525
Scene graphs are a compact and explicit representation successfully used in a va
riety of 2D scene understanding tasks. This work proposes a method to build up s
emantic scene graphs from a 3D environment incrementally given a sequence of RGB
-D frames. To this end, we aggregate PointNet features from primitive scene comp
onents by means of a graph neural network. We also propose a novel attention mec
hanism well suited for partial and missing graph data present in such an increme
ntal reconstruction scenario. Although our proposed method is designed to run on
 submaps of the scene, we show it also transfers to entire 3D scenes. Experiment
s show that our approach outperforms 3D scene graph prediction methods by a larg
e margin and its accuracy is on par with other 3D semantic and panoptic segmenta
tion methods while running at 35 Hz.
****
Interventional Video Grounding With Dual Contrastive Learning
Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, Wei Lu; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2021, pp. 2765-2775
Video grounding aims to localize a moment from an untrimmed video for a given te
xtual query. Existing approaches focus more on the alignment of visual and langu
age stimuli with various likelihood-based matching or regression strategies, i.e
., P(Y|X). Consequently, these models may suffer from spurious correlations betw
een the language and video features due to the selection bias of the dataset. 1)
 To uncover the causality behind the model and data, we first propose a novel pa
radigm from the perspective of the causal inference, i.e., interventional video
grounding (IVG) that leverages backdoor adjustment to deconfound the selection b
ias based on structured causal model (SCM) and do-calculus P(Y|do(X)). Then, we
present a simple yet effective method to approximate the unobserved confounder a
s it cannot be directly sampled from the dataset. 2) Meanwhile, we introduce a d
ual contrastive learning approach (DCL) to better align the text and video by ma
ximizing the mutual information (MI) between query and video clips, and the MI b
etween start/end frames of a target moment and the others within a video to lear
n more informative visual representations. Experiments on three standard benchma
rks show the effectiveness of our approaches.
****
A Fourier-Based Framework for Domain Generalization
Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, Qi Tian; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1
4383-14392
Modern deep neural networks suffer from performance degradation when evaluated o
n testing data under different distributions from training data. Domain generali
zation aims at tackling this problem by learning transferable knowledge from mul
tiple source domains in order to generalize to unseen target domains. This paper
 introduces a novel Fourier-based perspective for domain generalization. The mai
n assumption is that the Fourier phase information contains high-level semantics
 and is not easily affected by domain shifts. To force the model to capture phas
e information, we develop a novel Fourier-based data augmentation strategy calle
d amplitude mix which linearly interpolates between the amplitude spectrums of t
wo images. A dual-formed consistency loss called co-teacher regularization is fu
rther introduced between the predictions induced from original and augmented ima
ges. Extensive experiments on three benchmarks have demonstrated that the propos
ed method is able to achieve state-of-the-arts performance for domain generaliza
tion.
****
Probabilistic Modeling of Semantic Ambiguity for Scene Graph Generation
Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, Yujiu Yang; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
 pp. 12527-12536
To generate "accurate" scene graphs, almost all exist-ing methods predict pairwi
se relationships in a determin-istic manner. However, we argue that visual relat
ionshipsare often semantically ambiguous. Specifically, inspired bylinguistic kn

owledge, we classify the ambiguity into threetypes: Synonymy Ambiguity, Hyponymy Ambiguity, andMulti-view Ambiguity. The ambiguity naturally leads to theissue of implicit multi-label, motivating the need for diversepredictions. In this work, we propose a novel plug-and-play Probabilistic Uncertainty Modeling (PUM) module. Itmodels each union region as a Gaussian distribution, whosevariance measures the uncertainty of the corresponding vi-sual content. Compared to the conventional determinis-tic methods, such uncertainty modeling brings stochasticityof feature representation, which naturally enables diversepredictions. As a byproduct, PUM also manages to covermore fine-grained relationships and thus alleviates the is-sue of bias towards frequent relationships. Extensive exper-iments on the large-scale Visual Genome benchmark showthat combining PUM with newly proposed ResCAGCN canachieve state-of-the-art performances, especially under themean recall metric. Furthermore, we show the universal ef-fectiveness of PUM by plugging it into some existing modelsand provide insightful analysis of its ability to generate di-verse yet plausible visual relationships.
****
SRWarp: Generalized Image Super-Resolution under Arbitrary Transformation
Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7782-7791
Deep CNNs have achieved significant successes in image processing and its applications, including single image super-resolution (SR). However, conventional methods still resort to some predetermined integer scaling factors, e.g., x2 or x4. Thus, they are difficult to be applied when arbitrary target resolutions are required. Recent approaches extend the scope to real-valued upsampling factors, even with varying aspect ratios to handle the limitation. In this paper, we propose the SRWarp framework to further generalize the SR tasks toward an arbitrary image transformation. We interpret the traditional image warping task, specifically when the input is enlarged, as a spatially-varying SR problem. We also propose several novel formulations, including the adaptive warping layer and multiscale blending, to reconstruct visually favorable results in the transformation process. Compared with previous methods, we do not constrain the SR model on a regular grid but allow numerous possible deformations for flexible and diverse image editing. Extensive experiments and ablation studies justify the necessity and demonstrate the advantage of the proposed SRWarp method under various transformations.
****
IQDet: Instance-Wise Quality Distribution Sampling for Object Detection
Yuchen Ma, Songtao Liu, Zeming Li, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1717-1725
We propose a dense object detector with an instance-wise sampling strategy, named IQDet. Instead of using human prior sampling strategies, we first extract the regional feature of each ground-truth to estimate the instance-wise quality distribution. According to a mixture model in spatial dimensions, the distribution is more noise-robust and adapted to the semantic pattern of each instance. Based on the distribution, we propose a quality sampling strategy, which automatically selects training samples in a probabilistic manner and trains with more high-quality samples. Extensive experiments on MS COCO show that our method steadily improves baseline by nearly 2.4 AP without bells and whistles. Moreover, our best model achieves 51.6 AP, outperforming all existing state-of-the-art one-stage detectors and it is completely cost-free in inference time.
****
Scan2Cap: Context-Aware Dense Captioning in RGB-D Scans
Zhenyu Chen, Ali Gholami, Matthias Niessner, Angel X. Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3193-3203
We introduce the new task of dense captioning in RGB-D scans. As input, we assume a point cloud of a 3D scene; the expected output is the bounding boxes along with the descriptions for the underlying objects. To address the 3D object detecting and describing problem at the same time, we propose Scan2Cap, an end-to-end trained architecture, to detect objects in the input scene and generate the desc

riptions for all of them in natural language. We apply an attention-based captioning method to generate descriptive tokens while referring to the related components in the local context. To better handle the relative spatial relations between objects, a message passing graph module is applied to learn the relation features, which are later used in the captioning phase. On the recently proposed ScanRefer dataset, we show that our architecture can effectively localize and describe the 3D objects in the scene. It also outperforms the 2D-based methods on the 3D dense captioning task by a big margin.
****
NeuralHumanFVV: Real-Time Neural Volumetric Human Performance Rendering Using RGB Cameras
Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6226-6237
4D reconstruction and rendering of human activities is critical for immersive VR/AR experience. Recent advances still fail to recover fine geometry and texture results with the level of detail present in the input images from sparse multi-view RGB cameras. In this paper, we propose NeuralHumanFVV, a real-time neural human performance capture and rendering system to generate both high-quality geometry and photo-realistic texture of human activities in arbitrary novel views. We propose a neural geometry generation scheme with a hierarchical sampling strategy for real-time implicit geometry inference, as well as a novel neural blending scheme to generate high resolution (e.g., 1k) and photo-realistic texture results in the novel views. Furthermore, we adopt neural normal blending to enhance geometry details and formulate our neural geometry and texture rendering into a multi-task learning framework. Extensive experiments demonstrate the effectiveness of our approach to achieve high-quality geometry and photo-realistic free view-point reconstruction for challenging human performances.
****
Anti-Aliasing Semantic Reconstruction for Few-Shot Semantic Segmentation
Binghao Liu, Yao Ding, Jianbin Jiao, Xiangyang Ji, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9747-9756
Encouraging progress in few-shot semantic segmentation has been made by leveraging features learned upon base classes with sufficient training data to represent novel classes with few-shot examples. However, this feature sharing mechanism inevitably causes semantic aliasing between novel classes when they have similar compositions of semantic concepts. In this paper, we reformulate few-shot segmentation as a semantic reconstruction problem, and convert base class features into a series of basis vectors which span a class-level semantic space for novel class reconstruction. By introducing contrastive loss, we maximize the orthogonality of basis vectors while minimizing semantic aliasing between classes. Within the reconstructed representation space, we further suppress interference from other classes by projecting query features to the support vector for precise semantic activation. Our proposed approach, referred to as anti-aliasing semantic reconstruction (ASR), provides a systematic yet interpretable solution for few-shot learning problems. Extensive experiments on PASCAL VOC and MS COCO datasets show that ASR achieves strong results compared with the prior works. Code will be released at github.com/Bibkiller/ASR.
****
Composing Photos Like a Photographer
Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, Weicai Zhong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7057-7066
We show that explicit modeling of composition rules benefits image cropping. Image cropping is considered a promising way to automate aesthetic composition in professional photography. Existing efforts, however, only model such professional knowledge implicitly, e.g., by ranking from comparative candidates. Inspired by the observation that natural composition traits always follow a specific rule, we propose to learn such rules in a discriminative manner, and more importantly,

 to incorporate learned composition clues explicitly in the model. To this end, we introduce the concept of the key composition map (KCM) to encode the composition rules. The KCM can reveal the common laws hidden behind different composition rules and can inform the cropping model of what is important in composition. With the KCM, we present a novel cropping-by-composition paradigm and instantiate a network to implement composition-aware image cropping. Extensive experiments on two benchmarks justify that our approach enables effective, interpretable, and fast image cropping.
****
Asymmetric Gained Deep Image Compression With Continuous Rate Adaptation
Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, Bo Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10532-10541
With the development of deep learning techniques, the combination of deep learning with image compression has drawn lots of attention. Recently, learned image compression methods had exceeded their classical counterparts in terms of rate-distortion performance. However, continuous rate adaptation remains an open question. Some learned image compression methods use multiple networks for multiple rates, while others use one single model at the expense of computational complexity increase and performance degradation. In this paper, we propose a continuously rate adjustable learned image compression framework, Asymmetric Gained Variational Autoencoder (AG-VAE). AG-VAE utilizes a pair of gain units to achieve discrete rate adaptation in one single model with a negligible additional computation. Then, by using exponential interpolation, continuous rate adaptation is achieved without compromising performance. Besides, we propose the asymmetric Gaussian entropy model for more accurate entropy estimation. Exhaustive experiments show that our method achieves comparable quantitative performance with SOTA learned image compression methods and better qualitative performance than classical image codecs. In the ablation study, we confirm the usefulness and superiority of gain units and the asymmetric Gaussian entropy model.
****
Optimal Gradient Checkpoint Search for Arbitrary Computation Graphs
Jianwei Feng, Dong Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11433-11442
Deep Neural Networks(DNNs) require huge GPU memory when training on modern image/video databases. Unfortunately, the GPU memory is physically finite, which limits the image resolutions and batch sizes that could be used in training for better DNN performance. Unlike solutions that require physically upgrade GPUs, the Gradient CheckPointing(GCP) training trades computation for more memory beyond existing GPU hardware. GCP only stores a subset of intermediate tensors, called Gradient Checkpoints (GCs), during forward. Then during backward, extra local forwards are conducted to compute the missing tensors. The total training memory cost becomes the sum of (1) the memory cost of the gradient checkpoints and (2) the maximum memory cost of local forwards. To achieve maximal memory cut-offs, one needs optimal algorithms to select GCs. Existing GCP approaches rely on either manual input of GCs or heuristics-based GC search on Linear Computation Graphs (LCGs), and cannot apply to Arbitrary Computation Graphs(ACGs). In this paper, we present theories and optimal algorithms on GC selection that, for the first time, are applicable to ACGs and achieve the maximal memory cut-offs. Extensive experiments show that our approach not only outperforms existing approaches (only applicable on LCGs), and is applicable to a vast family of LCG and ACG networks, such as Alexnet, VGG, ResNet, Densenet, Inception Net and highly complicated DNNs by Network Architecture Search. Our work enables GCP training on ACGs, and cuts off up-to 80% of training memory with a moderate time overhead (30%-50%). Codes are available
****
NBNet: Noise Basis Learning for Image Denoising With Subspace Projection
Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4896-4906

In this paper, we introduce NBNet, a novel framework for image denoising. Unlike previous works, we propose to tackle this challenging problem from a new perspective: noise reduction by image-adaptive projection. Specifically, we propose to train a network that can separate signal and noise by learning a set of reconstruction basis in the feature space. Subsequently, image denosing can be achieved by selecting corresponding basis of the signal subspace and projecting the input into such space. Our key insight is that projection can naturally maintain the local structure of input signal, especially for areas with low light or weak textures. Towards this end, we propose SSA, a non-local attention module we design to explicitly learn the basis generation as well as subspace projection. We further incorporate SSA with NBNet, a UNet structured network designed for end-to-end image denosing based. We conduct evaluations on benchmarks, including SIDD and DND, and NBNet achieves state-of-the-art performance on PSNR and SSIM with significantly less computational cost.
****
NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis
Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, Jonathan T. Barron; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7495-7504
We present a method that takes as input a set of images of a scene illuminated by unconstrained known lighting, and produces as output a 3D representation that can be rendered from novel viewpoints under arbitrary lighting conditions. Our method represents the scene as a continuous volumetric function parameterized as MLPs whose inputs are a 3D location and whose outputs are the following scene properties at that input location: volume density, surface normal, material parameters, distance to the first surface intersection in any direction, and visibility of the external environment in any direction. Together, these allow us to render novel views of the object under arbitrary lighting, including indirect illumination effects. The predicted visibility and surface intersection fields are critical to our model's ability to simulate direct and indirect illumination during training, because the brute-force techniques used by prior work are intractable for lighting conditions outside of controlled setups with a single light. Our method outperforms alternative approaches for recovering relightable 3D scene representations, and performs well in complex lighting settings that have posed a significant challenge to prior work.
****
How Transferable Are Reasoning Patterns in VQA?
Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot, Christian Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4207-4216
Since its inception, Visual Question Answering (VQA) is notoriously known as a task, where models are prone to exploit biases in datasets to find shortcuts instead of performing high-level reasoning. Classical methods address this by removing biases from training data, or adding branches to models to detect and remove biases. In this paper, we argue that uncertainty in vision is a dominating factor preventing the successful learning of reasoning in vision and language problems. We train a visual oracle and in a large scale study provide experimental evidence that it is much less prone to exploiting spurious dataset biases compared to standard models. We propose to study the attention mechanisms at work in the visual oracle and compare them with a SOTA Transformer-based model. We provide an in-depth analysis and visualizations of reasoning patterns obtained with an online visualization tool which we make publicly available (https://reasoningpatterns.github.io). We exploit these insights by transferring reasoning patterns from the oracle to a SOTA Transformer-based VQA model taking standard noisy visual inputs via fine-tuning. In experiments we report higher overall accuracy, as well as accuracy on infrequent answers for each question type, which provides evidence for improved generalization and a decrease of the dependency on dataset biases.
****
DyStaB: Unsupervised Object Segmentation via Dynamic-Static Bootstrapping

Yanchao Yang, Brian Lai, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2826-2836
We describe an unsupervised method to detect and segment portions of images of live scenes that, at some point in time, are seen moving as a coherent whole, which we refer to as objects. Our method first partitions the motion field by minimizing the mutual information between segments. Then, it uses the segments to learn object models that can be used for detection in a static image. Static and dynamic models are represented by deep neural networks trained jointly in a bootstrapping strategy, which enables extrapolation to previously unseen objects. While the training process requires motion, the resulting object segmentation network can be used on either static images or videos at inference time. As the volume of seen videos grows, more and more objects are seen moving, priming their detection, which then serves as a regularizer for new objects, turning our method into unsupervised continual learning to segment objects. Our models are compared to the state of the art in both video object segmentation and salient object detection. In the six benchmark datasets tested, our models compare favorably even to those using pixel-level supervision, despite requiring no manual annotation.
****
Deep Texture Recognition via Exploiting Cross-Layer Statistical Self-Similarity
Zhile Chen, Feng Li, Yuhui Quan, Yong Xu, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5231-5240
In recent years, convolutional neural networks (CNNs) have become a prominent tool for texture recognition. The key of existing CNN-based approaches is aggregating the convolutional features into a robust yet discriminative description. This paper presents a novel feature aggregation module called CLASS (Cross-Layer Aggregation of Statistical Self-similarity) for texture recognition. We model the CNN feature maps across different layers, as a dynamic process which carries the statistical self-similarity (SSS), one well-known property of texture, from input image along the network depth dimension. The CLASS module characterizes the cross-layer SSS using a soft histogram of local differential box-counting dimensions of cross-layer features. The resulting descriptor encodes both cross-layer dynamics and local SSS of input image, providing additional discrimination over the often-used global average pooling. Integrating CLASS into a ResNet backbone, we develop CLASSNet, an effective deep model for texture recognition, which shows state-of-the-art performance in the experiments.
****
Light Field Super-Resolution With Zero-Shot Learning
Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10010-10019
Deep learning provides a new avenue for light field super-resolution (SR). However, the domain gap caused by drastically different light field acquisition conditions poses a main obstacle in practice. To fill this gap, we propose a zero-shot learning framework for light field SR, which learns a mapping to super-resolve the reference view with examples extracted solely from the input low-resolution light field itself. Given highly limited training data under the zero-shot setting, however, we observe that it is difficult to train an end-to-end network successfully. Instead, we divide this challenging task into three sub-tasks, i.e., pre-upsampling, view alignment, and multi-view aggregation, and then conquer them separately with simple yet efficient CNNs. Moreover, the proposed framework can be readily extended to finetune the pre-trained model on a source dataset to better adapt to the target input, which further boosts the performance of light field SR in the wild. Experimental results validate that our method not only outperforms classic non-learning-based methods, but also generalizes better to unseen light fields than state-of-the-art deep-learning-based methods when the domain gap is large.
****
Spherical Confidence Learning for Face Recognition
Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, Bryan Hooi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP

R), 2021, pp. 15629-15637
An emerging line of research has found that spherical spaces better match the un
derlying geometry of facial images, as evidenced by the state-of-the-art facial
recognition methods which benefit empirically from spherical representations. Ye
t, these approaches rely on deterministic embeddings and hence suffer from the f
eature ambiguity dilemma, whereby ambiguous or noisy images are mapped into poor
ly learned regions of representation space, leading to inaccuracies. Probabilist
ic Face Embeddings (PFE) is the first attempt to address this dilemma. However,
we theoretically and empirically identify two main failures of PFE when it is ap
plied to spherical deterministic embeddings aforementioned. To address these iss
ues, in this paper, we propose a novel framework for face confidence learning in
 spherical space. Mathematically, we extend the von Mises Fisher density to its
r-radius counterpart and derive a new optimization objective in closed form. The
oretically, the proposed probabilistic framework provably allows for better inte
rpretability, leading to principled feature comparison and pooling. Extensive ex
perimental results on multiple challenging benchmarks confirm our hypothesis and
 theory, and showcase the advantages of our framework over prior probabilistic m
ethods and spherical deterministic embeddings in various face recognition tasks.
****
Three Ways To Improve Semantic Segmentation With Self-Supervised Depth Estimatio
n
Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, Luc Van Gool; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2021, pp. 11130-11140
Training deep networks for semantic segmentation requires large amounts of label
ed training data, which presents a major challenge in practice, as labeling segm
entation masks is a highly labor-intensive process. To address this issue, we pr
esent a framework for semi-supervised semantic segmentation, which is enhanced b
y self-supervised monocular depth estimation from unlabeled image sequences. In
particular, we propose three key contributions: (1) We transfer knowledge from f
eatures learned during self-supervised depth estimation to semantic segmentation
, (2) we implement a strong data augmentation by blending images and labels usin
g the geometry of the scene, and (3) we utilize the depth feature diversity as w
ell as the level of difficulty of learning depth in a student-teacher framework
to select the most useful samples to be annotated for semantic segmentation. We
validate the proposed model on the Cityscapes dataset, where all three modules d
emonstrate significant performance gains, and we achieve state-of-the-art result
s for semi-supervised semantic segmentation. The implementation is available at
https://github.com/lhoyer/improving_segmentation_with_selfsupervised_depth.
****
Cross-Modal Contrastive Learning for Text-to-Image Generation
Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, Yinfei Yang; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 833-842
The output of text-to-image synthesis systems should be coherent, clear, photo-r
ealistic scenes with high semantic fidelity to their conditioned text descriptio
ns. Our Cross-Modal Contrastive Generative Adversarial Network (XMC-GAN) address
es this challenge by maximizing the mutual information between image and text. I
t does this via multiple contrastive losses which capture inter-modality and int
ra-modality correspondences. XMC-GAN uses an attentional self-modulation generat
or, which enforces strong text-image correspondence, and a contrastive discrimin
ator, which acts as a critic as well as a feature encoder for contrastive learni
ng. The quality of XMC-GAN's output is a major step up from previous models, as
we show on three challenging datasets. On MS-COCO, not only does XMC-GAN improve
 state-of-the-art FID from 24.70 to 9.33, but--more importantly--people prefer X
MC-GAN by 77.3 for image quality and 74.1 for image-text alignment, compared to
three other recent models. XMC-GAN also generalizes to the challenging Localized
 Narratives dataset (which has longer, more detailed descriptions), improving st
ate-of-the-art FID from 48.70 to 14.12. Lastly, we train and evaluate XMC-GAN on
 the challenging Open Images data, establishing a strong benchmark FID score of

26.91.
****
Lifting 2D StyleGAN for 3D-Aware Face Generation
Yichun Shi, Divyansh Aggarwal, Anil K. Jain; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6258-6266
We propose a framework, called LiftedGAN, that disentangles and lifts a pre-trained StyleGAN2 for 3D-aware face generation. Our model is "3D-aware" in the sense that it is able to (1) disentangle the latent space of StyleGAN2 into texture, shape, viewpoint, lighting and (2) generate 3D components for rendering synthetic images. Unlike most previous methods, our method is completely self-supervised, i.e. it neither requires any manual annotation nor 3DMM model for training. Instead, it learns to generate images as well as their 3D components by distilling the prior knowledge in StyleGAN2 with a differentiable renderer. The proposed model is able to output both the 3D shape and texture, allowing explicit pose and lighting control over generated images. Qualitative and quantitative results show the superiority of our approach over existing methods on 3D-controllable GANs in content controllability while generating realistic high quality images.
****
iMiGUE: An Identity-Free Video Dataset for Micro-Gesture Understanding and Emotion Analysis
Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, Guoying Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10631-10642
We introduce a new dataset for the emotional artificial intelligence research: identity-free video dataset for micro-gesture understanding and emotion analysis (iMiGUE). Different from existing public datasets, iMiGUE focuses on nonverbal body gestures without using any identity information, while the predominant researches of emotion analysis concern sensitive biometric data, like face and speech. Most importantly, iMiGUE focuses on micro-gestures, i,e., unintentional behaviors driven by inner feelings, which are different from ordinary scope of gestures from other gesture datasets which are mostly intentionally performed for illustrative purposes. Furthermore, iMiGUE is designed to evaluate the ability of models to analyze the emotional states by integrating information of recognized micro-gesture, rather than just recognizing prototypes in the sequences separately (or isolatedly). This is because the real need for emotion AI is to understand the emotional states behind gestures in a holistic way. Moreover, to counter for the challenge of imbalanced samples distribution of this dataset, an unsupervised learning method is proposed to capture latent representations from the micro-gesture sequences themselves. We systematically investigate representative methods on this dataset, and comprehensive experimental results reveal several interesting insights from the iMiGUE, e,g., micro-gesture-based analysis can promote emotion understanding. We confirm that the new iMiGUE dataset could advance studies of micro-gesture and emotion AI.
****
MeGA-CDA: Memory Guided Attention for Category-Aware Unsupervised Domain Adaptive Object Detection
Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4516-4526
Existing approaches for unsupervised domain adaptive object detection perform feature alignment via adversarial training. While these methods achieve reasonable improvements in performance, they typically perform category-agnostic domain alignment, thereby resulting in negative transfer of features. To overcome this issue, in this work, we attempt to incorporate category information into the domain adaptation process by proposing Memory Guided Attention for Category-Aware Domain Adaptation (MeGA-CDA). The proposed method consists of employing category-wise discriminators to ensure category-aware feature alignment for learning domain-invariant discriminative features. However, since the category information is not available for the target samples, we propose to generate memory-guided category-specific attention maps which are then used to route the features appropriate

ly to the corresponding category discriminator. The proposed method is evaluated on several benchmark datasets and is shown to outperform existing approaches.
****
Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food
Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, Jack Sim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8903-8911
Understanding the nutritional content of food from visual data is a challenging computer vision problem, with the potential to have a positive and widespread impact on public health. Studies in this area are limited to existing datasets in the field that lack sufficient diversity or labels required for training models with nutritional understanding capability. We introduce Nutrition5k, a novel dataset of 5k diverse, real world food dishes with corresponding video streams, depth images, component weights, and high accuracy nutritional content annotation. We demonstrate the potential of this dataset by training a computer vision algorithm capable of predicting the caloric and macronutrient values of a complex, real world dish at an accuracy that outperforms professional nutritionists. Further we present a baseline for incorporating depth sensor data to improve nutrition predictions. We release Nutrition5k in the hope that it will accelerate innovation in the space of nutritional understanding. The dataset is available at https://github.com/google-research-datasets/Nutrition5k.
****
Extreme Low-Light Environment-Driven Image Denoising Over Permanently Shadowed Lunar Regions With a Physical Noise Model
Ben Moseley, Valentin Bickel, Ignacio G. Lopez-Francos, Loveneesh Rana; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6317-6327
Recently, learning-based approaches have achieved impressive results in the field of low-light image denoising. Some state of the art approaches employ a rich physical model to generate realistic training data. However, the performance of these approaches ultimately depends on the realism of the physical model, and many works only concentrate on everyday photography. In this work we present a denoising approach for extremely low-light images of permanently shadowed regions (PSRs) on the lunar surface, taken by the Narrow Angle Camera on board the Lunar Reconnaissance Orbiter satellite. Our approach extends existing learning-based approaches by combining a physical noise model of the camera with real noise samples and training image scene selection based on 3D ray tracing to generate realistic training data. We also condition our denoising model on the camera's environmental metadata at the time of image capture (such as the camera's temperature and age), showing that this improves performance. Our quantitative and qualitative results show that our method strongly outperforms the existing calibration routine for the camera and other baselines. Our results could significantly impact lunar science and exploration, for example by aiding the identification of surface water-ice and reducing uncertainty in rover and human traverse planning into PSRs.
****
Unsupervised Discovery of the Long-Tail in Instance Segmentation Using Hierarchical Self-Supervision
Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, Serena Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2603-2612
Instance segmentation is an active topic in computer vision that is usually solved by using supervised learning approaches over very large datasets composed of object level masks. Obtaining such a dataset for any new domain can be very expensive and time-consuming. In addition, models trained on certain annotated categories do not generalize well to unseen objects. The goal of this paper is to propose a method that can perform unsupervised discovery of long-tail categories in instance segmentation, through learning instance embeddings of masked regions. Leveraging rich relationship and hierarchical structure between objects in the images, we propose self-supervised losses for learning mask embeddings. Trained o

n COCO dataset without additional annotations of the long-tail objects, our mode l is able to discover novel and more fine-grained objects than the common catego ries in COCO. We show that the model achieves competitive quantitative results o n LVIS as compared to the supervised and partially supervised methods.
****
How Privacy-Preserving Are Line Clouds? Recovering Scene Details From 3D Lines
Kunal Chelani, Fredrik Kahl, Torsten Sattler; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15668-15678
Visual localization is the problem of estimating the camera pose of a given imag e with respect to a known scene. Visual localization algorithms are a fundamenta l building block in advanced computer vision applications, including Mixed and V irtual Reality systems. Many algorithms used in practice represent the scene thr ough a Structure-from-Motion (SfM) point cloud, where each 3D point is associate d with one or more local image features. Establishing 2D-3D matches between feat ures in a query image and the 3D points through descriptor matching Visual local ization is the problem of estimating the camera pose of a given image with respe ct to a known scene. Visual localization algorithms are a fundamental building b lock in advanced computer vision applications, including Mixed and Virtual Reali ty systems. Many algorithms used in practice represent the scene through a Struc ture-from-Motion (SfM) point cloud and use 2D-3D matches between a query image a nd the 3D points for camera pose estimation. As recently shown, image details ca n be accurately recovered from SfM point clouds by translating renderings of the sparse point clouds to images. To address the resulting potential privacy risks for user-generated content, it was recently proposed to lift point clouds to li ne clouds by replacing 3D points by randomly oriented 3D lines passing through t hese points. The resulting representation is unintelligible to humans and effect ively prevents point cloud-to-image translation. This paper shows that a signifi cant amount of information about the 3D scene geometry is preserved in these lin e clouds, allowing us to (approximately) recover the 3D point positions and thus to (approximately) recover image content. Our approach is based on the observat ion that the closest points between lines can yield a good approximation to the original 3D points. Code is available at \href https://github.com/kunalchelani/L ine2Point https://github.com/kunalchelani/Line2Point .
****
Multi-View 3D Reconstruction of a Texture-Less Smooth Surface of Unknown Generic Reflectance
Ziang Cheng, Hongdong Li, Yuta Asano, Yinqiang Zheng, Imari Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 , pp. 16226-16235
Recovering the 3D geometry of a purely texture-less object with generally unknow n surface reflectance (e.g. nonLambertian) is regarded as a challenging task in multiview reconstruction. The major obstacle revolves around establishing cross-view correspondences where photometric constancy is violated. This paper propose s a simple and practical solution to overcome this challenge based on a co-locat ed camera-light scanner device. Unlike existing solutions, we do not explicitly solve for correspondence. Instead, we argue the problem is generally well-posed by multi-view geometrical and photometric constraints, and can be solved from a small number of input views. We formulate the reconstruction task as a joint ene rgy minimization over the surface geometry and reflectance. Despite this energy is highly non-convex, we develop an optimization algorithm that robustly recover s globally optimal shape and reflectance even from a random initialization. Exte nsive experiments on both simulated and real data have validated our method, and possible future extensions are discussed
****
Rectification-Based Knowledge Retention for Continual Learning
Pravendra Singh, Pratik Mazumder, Piyush Rai, Vinay P. Namboodiri; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 1, pp. 15282-15291
Deep learning models suffer from catastrophic forgetting when trained in an incr emental learning setting. In this work, we propose a novel approach to address t

he task incremental learning problem, which involves training a model on new tas ks that arrive in an incremental manner. The task incremental learning problem b ecomes even more challenging when the test set contains classes that are not par t of the train set, i.e., a task incremental generalized zero-shot learning prob lem. Our approach can be used in both the zero-shot and non zero-shot task incre mental learning settings. Our proposed method uses weight rectifications and aff ine transformations in order to adapt the model to different tasks that arrive s equentially. Specifically, we adapt the network weights to work for new tasks by "rectifying" the weights learned from the previous task. We learn these weight rectifications using very few parameters. We additionally learn affine transform ations on the outputs generated by the network in order to better adapt them for the new task. We perform experiments on several datasets in both zero-shot and non zero-shot task incremental learning settings and empirically show that our a pproach achieves state-of-the-art results. Specifically, our approach outperform s the state-of-the-art non zero-shot task incremental learning method by over 5% on the CIFAR-100 dataset. Our approach also significantly outperforms the state -of-the-art task incremental generalized zero-shot learning method by absolute m argins of 6.91% and 6.33% for the AWA1 and CUB datasets, respectively. We valida te our approach using various ablation studies.
****
Scale-Aware Automatic Augmentation for Object Detection
Yukang Chen, Yanwei Li, Tao Kong, Lu Qi, Ruihang Chu, Lei Li, Jiaya Jia; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9563-9572
We propose Scale-aware AutoAug to learn data augmentation policies for object de tection. We define a new scale-aware search space, where both image- and box-lev el augmentations are designed for maintaining scale invariance. Upon this search space, we propose a new search metric, termed Pareto Scale Balance, to facilita te search with high efficiency. In experiments, Scale-aware AutoAug yields signi ficant and consistent improvement on various object detectors (e.g., RetinaNet, Faster R-CNN, Mask R-CNN, and FCOS), even compared with strong multi-scale train ing baselines. Our searched augmentation policies are transferable to other data sets and box-level tasks beyond object detection (e.g., instance segmentation an d keypoint estimation) to improve performance. The search cost is much less than previous automated augmentation approaches for object detection. It is notable that our searched policies have meaningful patterns, which intuitively provide v aluable insight for human data augmentation design.
****
Towards Robust Classification Model by Counterfactual and Invariant Data Generat ion
Chun-Hao Chang, George Alexandru Adam, Anna Goldenberg; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1521 2-15221
Despite the success of machine learning applications in science, industry, and s ociety in general, many approaches are known to be non-robust, often relying on spurious correlations to make predictions. Spuriousness occurs when some feature s correlate with labels but are not causal; relying on such features prevents mo dels from generalizing to unseen environments where such correlations break. In this work, we focus on image classification and propose two data generation proc esses to reduce spuriousness. Given human annotations of the subset of the featu res responsible (causal) for the labels (e.g. bounding boxes), we modify this ca usal set to generate a surrogate image that no longer has the same label (i.e. a counterfactual image). We also alter non-causal features to generate images sti ll recognized as the original labels, which helps to learn a model invariant to these features. In several challenging datasets, our data generations outperform state-of-the-art methods in accuracy when spurious correlations break, and incr ease the saliency focus on causal features providing better explanations.
****
Fully Convolutional Networks for Panoptic Segmentation
Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, Jiaya

Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2021, pp. 214-223
In this paper, we present a conceptually simple, strong, and efficient framework for panoptic segmentation, called Panoptic FCN. Our approach aims to represent and predict foreground things and background stuff in a unified fully convolutional pipeline. In particular, Panoptic FCN encodes each object instance or stuff category into a specific kernel weight with the proposed kernel generator and produces the prediction by convolving the high-resolution feature directly. With this approach, instance-aware and semantically consistent properties for things and stuff can be respectively satisfied in a simple generate-kernel-then-segment workflow. Without extra boxes for localization or instance separation, the proposed approach outperforms previous box-based and -free models with high efficiency on COCO, Cityscapes, and Mapillary Vistas datasets with single scale input. Our code is made publicly available at https://github.com/Jia-Research-Lab/PanopticFCN.
****
Benchmarking Representation Learning for Natural World Image Collections
Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, Oisin Mac Aodha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12884-12893
Recent progress in self-supervised learning has resulted in models that are capable of extracting rich representations from image collections without requiring any explicit label supervision. However, to date the vast majority of these approaches have restricted themselves to training on standard benchmark datasets such as ImageNet. We argue that fine-grained visual categorization problems, such as plant and animal species classification, provide an informative testbed for self-supervised learning. In order to facilitate progress in this area we present two new natural world visual classification datasets, iNat2021 and NeWT. The former consists of 2.7M images from 10k different species uploaded by users of the citizen science application iNaturalist. We designed the latter, NeWT, in collaboration with domain experts with the aim of benchmarking the performance of representation learning algorithms on a suite of challenging natural world binary classification tasks that go beyond standard species classification. These two new datasets allow us to explore questions related to large-scale representation and transfer learning in the context of fine-grained categories. We provide a comprehensive analysis of feature extractors trained with and without supervision on ImageNet and iNat2021, shedding light on the strengths and weaknesses of different learned features across a diverse set of tasks. We find that features produced by standard supervised methods still outperform those produced by self-supervised approaches such as SimCLR. However, improved self-supervised learning methods are constantly being released and the iNat2021 and NeWT datasets are a valuable resource for tracking their progress.
****
PGT: A Progressive Method for Training Models on Long Videos
Bo Pang, Gao Peng, Yizhuo Li, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11379-11389
Convolutional video models have an order of magnitude larger computational complexity than their counterpart image-level models. Constrained by computational resources, there is no model or training method that can train long video sequences end-to-end. Currently, the main-stream method is to split a raw video into clips, leading to incomplete fragmentary temporal information flow. Inspired by natural language processing techniques dealing with long sentences, we propose to treat videos as serial fragments satisfying Markov property, and train it as a whole by progressively propagating information through the temporal dimension in multiple steps. This progressive training (PGT) method is able to train long videos end-to-end with limited resources and ensures the effective transmission of information. As a general and robust training method, we empirically demonstrate that it yields significant performance improvements on different models and datasets. As an illustrative example, the proposed method improves SlowOnly network by 3.7 mAP on Charades and 1.9 top-1 accuracy on Kinetics with negligible parame

ter and computation overhead. The code is attached in supplementary files and will be published with this paper.
****
Prioritized Architecture Sampling With Monto-Carlo Tree Search
Xiu Su, Tao Huang, Yanxi Li, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10968-10977
One-shot neural architecture search (NAS) methods significantly reduce the search cost by considering the whole search space as one network, which only needs to be trained once. However, current methods select each operation independently without considering previous layers. Besides, the historical information obtained with huge computation costs is usually used only once and then discarded. In this paper, we introduce a sampling strategy based on Monte Carlo tree search (MCTS) with the search space modeled as a Monte Carlo tree (MCT), which captures the dependency among layers. Furthermore, intermediate results are stored in the MCT for future decisions and a better exploration-exploitation balance. Concretely, MCT is updated using the training loss as a reward to the architecture performance; for accurately evaluating the numerous nodes, we propose node communication and hierarchical node selection methods in the training and search stages, respectively, making better uses of the operation rewards and hierarchical information. Moreover, for a fair comparison of different NAS methods, we construct an open-source NAS benchmark of a macro search space evaluated on CIFAR-10, namely NAS-Bench-Macro. Extensive experiments on NAS-Bench-Macro and ImageNet demonstrate that our method significantly improves search efficiency and performance. For example, by only searching 20 architectures, our obtained architecture achieves 78.0% top-1 accuracy with 442M FLOPs on ImageNet. Code (Benchmark) is available at: https://github.com/xiusu/NAS-Bench-Macro.
****
HumanGPS: Geodesic PreServing Feature for Dense Human Correspondences
Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, Ping Tan, Yinda Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1820-1830
In this paper, we address the problem of building pixel-wise dense correspondences between human images under arbitrary camera viewpoints and body poses. Previous methods either assume small motions or rely on discriminative descriptors extracted from local patches, which cannot handle large motion or visually ambiguous body parts, e.g. left v.s. right hand. In contrast, we propose a deep learning framework that maps each pixel to a feature space, where the feature distances reflect the geodesic distances among pixels as if they were projected onto the surface of 3D human scans. To this end, we introduce novel loss functions to push features apart according to their geodesic distances on the surface inside and across images. Without any semantic annotation, the features automatically learn to differentiate visually similar parts and align different subjects into a unified feature space. Extensive experiments show that the learned features can produce accurate correspondences between images with remarkable generalization capabilities on both intra and inter subjects. We demonstrate the effectiveness of our method on a variety of applications such as optical flow, non-rigid tracking, occlusions detection, and human dense pose regression.
****
Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition
Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7098-7107
Linguistic knowledge is of great benefit to scene text recognition. However, how to effectively model linguistic rules in end-to-end deep networks remains a research challenge. In this paper, we argue that the limited capacity of language models comes from: 1) implicitly language modeling; 2) unidirectional feature representation; and 3) language model with noise input. Correspondingly, we propose

an autonomous, bidirectional and iterative ABINet for scene text recognition. Firstly, the autonomous suggests to block gradient flow between vision and language models to enforce explicitly language modeling. Secondly, a novel bidirectional cloze network (BCN) as the language model is proposed based on bidirectional feature representation. Thirdly, we propose an execution manner of iterative correction for language model which can effectively alleviate the impact of noise input. Additionally, based on the ensemble of iterative predictions, we propose a self-training method which can learn from unlabeled images effectively. Extensive experiments indicate that ABINet has superiority on low-quality images and achieves state-of-the-art results on several mainstream benchmarks. Besides, the ABINet trained with ensemble self-training shows promising improvement in realizing human-level recognition.
****
Generic Perceptual Loss for Modeling Structured Output Dependencies
Yifan Liu, Hao Chen, Yu Chen, Wei Yin, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5424-5432
The perceptual loss has been widely used as an effective loss term in image synthesis tasks including image super-resolution [16], and style transfer [14]. It was believed that the success lies in the high-level perceptual feature representations extracted from CNNs pretrained with a large set of images. Here we reveal that what matters is the network structure instead of the trained weights. Without any learning, the structure of a deep network is sufficient to capture the dependencies between multiple levels of variable statistics using multiple layers of CNNs. This insight removes the requirements of pre-training and a particular network structure (commonly, VGG) that are previously assumed for the perceptual loss, thus enabling a significantly wider range of applications. To this end, we demonstrate that a randomly-weighted deep CNN can be used to model the structured dependencies of outputs. On a few dense per-pixel prediction tasks such as semantic segmentation, depth estimation, and instance segmentation, we show improved results of using the extended randomized perceptual loss, compared to the baselines using pixel-wise loss alone. We hope that this simple, extended perceptual loss may serve as a generic structured-output loss that is applicable to most structured output learning tasks.
****
Style-Based Point Generator With Adversarial Rendering for Point Cloud Completion
Chulin Xie, Chuxin Wang, Bo Zhang, Hao Yang, Dong Chen, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4619-4628
In this paper, we proposed a novel Style-based Point Generator with Adversarial Rendering (SpareNet) for point cloud completion. Firstly, we present the channel-attentive EdgeConv to fully exploit the local structures as well as the global shape in point features. Secondly, we observe that the concatenation manner used by vanilla foldings limits its potential of generating a complex and faithful shape. Enlightened by the success of StyleGAN, we regard the shape feature as style code that modulates the normalization layers during the folding, which considerably enhances its capability. Thirdly, we realize that existing point supervisions, e.g., Chamfer Distance or Earth Mover's Distance, cannot faithfully reflect the perceptual quality of the reconstructed points. To address this, we propose to project the completed points to depth maps with a differentiable renderer and apply adversarial training to advocate the perceptual realism under different viewpoints. Comprehensive experiments on ShapeNet and KITTI prove the effectiveness of our method, which achieves state-of-the-art quantitative performance while offering superior visual quality.
****
Neural Architecture Search With Random Labels
Xuanyang Zhang, Pengfei Hou, Xiangyu Zhang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10907-10916

In this paper, we investigate a new variant of neural architecture search (NAS) paradigm -- searching with random labels (RLNAS). The task sounds counter-intuitive for most existing NAS algorithms since random label provides few information on the performance of each candidate architecture. Instead, we propose a novel NAS framework based on ease-of-convergence hypothesis, which requires only random labels during searching. The algorithm involves two steps: first, we train a SuperNet using random labels; second, from the SuperNet we extract the sub-network whose weights change most significantly during the training. Extensive experiments are evaluated on multiple datasets (e.g. NAS-Bench-201 and ImageNet) and multiple search spaces (e.g. DARTS-like and MobileNet-like). Very surprisingly, RLNAS achieves comparable or even better results compared with state-of-the-art NAS methods such as PC-DARTS, Single Path One-Shot, even though the counterparts utilize full ground truth labels for searching. We hope our finding could inspire new understandings on the essential of NAS.
****

Towards Long-Form Video Understanding
Chao-Yuan Wu, Philipp Krahenbuhl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1884-1894
Our world offers a never-ending stream of visual stimuli, yet today's vision systems only accurately recognize patterns within a few seconds. These systems understand the present, but fail to contextualize it in past or future events. In this paper, we study long-form video understanding. We introduce a framework for modeling long-form videos and develop evaluation protocols on large-scale datasets. We show that existing state-of-the-art short-term models are limited for long-form tasks. A novel object-centric transformer-based video recognition architecture performs significantly better on 7 diverse tasks. It also outperforms comparable state-of-the-art on the AVA dataset.
****

Shape and Material Capture at Home
Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, David W. Jacobs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6123-6133
In this paper, we present a technique for estimating the geometry and reflectance of objects using only a camera, flashlight, and optionally a tripod. We propose a simple data capture technique in which the user goes around the object, illuminating it with a flashlight and capturing only a few images. Our main technical contribution is the introduction of a recursive neural architecture, which can predict geometry and reflectance at 2^kx2^k resolution given an input image at 2^kx2^k and estimated geometry and reflectance from the previous step at 2^(k-1)x2^(k-1). This recursive architecture, termed RecNet, is trained with 256x256 resolution but can easily operate on 1024x1024 images during inference. We show that our method produces more accurate surface normal and albedo, especially in regions of specular highlights and cast shadows, compared to previous approaches, given three or fewer input images.
****

Deep Polarization Imaging for 3D Shape and SVBRDF Acquisition
Valentin Deschaintre, Yiming Lin, Abhijeet Ghosh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15567-15576
We present a novel method for efficient acquisition of shape and spatially varying reflectance of 3D objects using polarization cues. Unlike previous works that have exploited polarization to estimate material or object appearance under certain constraints (known shape or multiview acquisition), we lift such restrictions by coupling polarization imaging with deep learning to achieve high quality estimate of 3D object shape (surface normals and depth) and SVBRDF using single-view polarization imaging under frontal flash illumination. In addition to acquired polarization images, we provide our deep network with strong novel cues related to shape and reflectance, in the form of a normalized Stokes map and an estimate of diffuse color. We additionally describe modifications to network architecture and training loss which provide further qualitative improvements. We demons

trate our approach to achieve superior results compared to recent works employin
g deep learning in conjunction with flash illumination.
****
Convolutional Neural Network Pruning With Structural Redundancy Reduction
Zi Wang, Chengcheng Li, Xiangyang Wang; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14913-14922
Convolutional neural network (CNN) pruning has become one of the most successful
 network compression approaches in recent years. Existing works on network pruni
ng usually focus on removing the least important filters in the network to achie
ve compact architectures. In this study, we claim that identifying structural re
dundancy plays a more essential role than finding unimportant filters, theoretic
ally and empirically. We first statistically model the network pruning problem i
n a redundancy reduction perspective and find that pruning in the layer(s) with
the most structural redundancy outperforms pruning the least important filters a
cross all layers. Based on this finding, we then propose a network pruning appro
ach that identifies structural redundancy of a CNN and prunes filers in the sele
cted layer(s) with the most redundancy. Experiments on various benchmark network
 architectures and datasets show that our proposed approach significantly outper
forms the previous state-of-the-art.
****
T-vMF Similarity for Regularizing Intra-Class Feature Distribution
Takumi Kobayashi; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2021, pp. 6616-6625
Deep convolutional neural networks (CNNs) leverage large-scale training dataset
to produce remarkable performance on various image classification tasks. It, how
ever, is difficult to effectively train the CNNs on some realistic learning situ
ations such as regarding class imbalance, small-scale and label noises. Regulari
zing CNNs works well on learning with such deteriorated training datasets by mit
igating overfitting issues. In this work, we propose a method to effectively imp
ose regularization on feature representation learning. By focusing on the angle
between a feature and a classifier which is embedded in cosine similarity at the
 classification layer, we formulate a novel similarity beyond the cosine based o
n von Mises-Fisher distribution of directional statistics. In contrast to the co
sine similarity, our similarity is compact while having heavy tail, which contri
butes to regularizing intra-class feature distribution to improve generalization
 performance. Through the experiments on some realistic learning situations such
 as of imbalance, small-scale and noisy labels, we demonstrate the effectiveness
 of the proposed method for training CNNs, in comparison to the other regulariza
tion methods. Codes are available at https://github.com/tk1980/tvMF.
****
Surrogate Gradient Field for Latent Space Manipulation
Minjun Li, Yanghua Jin, Huachun Zhu; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2021, pp. 6529-6538
Generative adversarial networks (GANs) can generate high-quality images from sam
pled latent codes. Recent works attempt to edit an image by manipulating its und
erlying latent code, but rarely go beyond the basic task of attribute adjustment
. We propose the first method that enables manipulation with multidimensional co
ndition such as keypoints and captions. Specifically, we design an algorithm tha
t searches for a new latent code that satisfies the target condition based on th
e Surrogate Gradient Field (SGF) induced by an auxiliary mapping network. For qu
antitative comparison, we propose a metric to evaluate the disentanglement of ma
nipulation methods. Thorough experimental analysis on the facial attribute adjus
tment task shows that our method outperforms state-of-the-art methods in disenta
nglement. We further apply our method to tasks of various condition modalities t
o demonstrate that our method can alter complex image properties such as keypoin
ts and captions.
****
SCF-Net: Learning Spatial Contextual Features for Large-Scale Point Cloud Segmen
tation
Siqi Fan, Qiulei Dong, Fenghua Zhu, Yisheng Lv, Peijun Ye, Fei-Yue Wang; Proceed

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14504-14513
How to learn effective features from large-scale point clouds for semantic segmentation has attracted increasing attention in recent years. Addressing this problem, we propose a learnable module that learns Spatial Contextual Features from large-scale point clouds, called SCF in this paper. The proposed module mainly consists of three blocks, including the local polar representation block, the dual-distance attentive pooling block, and the global contextual feature block. For each 3D point, the local polar representation block is firstly explored to construct a spatial representation that is invariant to the z-axis rotation, then the dual-distance attentive pooling block is designed to utilize the representations of its neighbors for learning more discriminative local features according to both the geometric and feature distances among them, and finally, the global contextual feature block is designed to learn a global context for each 3D point by utilizing its spatial location and the volume ratio of the neighborhood to the global point cloud. The proposed module could be easily embedded into various network architectures for point cloud segmentation, naturally resulting in a new 3D semantic segmentation network with an encoder-decoder architecture, called SCF-Net in this work. Extensive experimental results on two public datasets demonstrate that the proposed SCF-Net performs better than several state-of-the-art methods in most cases.
****
UnsupervisedR&R: Unsupervised Point Cloud Registration via Differentiable Rendering
Mohamed El Banani, Luya Gao, Justin Johnson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7129-7139
Aligning partial views of a scene into a single whole is essential to understanding one's environment and is a key component of numerous robotics tasks such as SLAM and SfM. Recent approaches have proposed end-to-end systems that can outperform traditional methods by leveraging pose supervision. However, with the rising prevalence of cameras with depth sensors, we can expect a new stream of raw RGB-D data without the annotations needed for supervision. We propose UnsupervisedR&R: an end-to-end unsupervised approach to learning point cloud registration from raw RGB-D video. The key idea is to leverage differentiable alignment and rendering to enforce photometric and geometric consistency between frames. We evaluate our approach on indoor scene datasets and find that we outperform existing traditional approaches with classical and learned descriptors while being competitive with supervised geometric point cloud registration approaches.
****
ZeroScatter: Domain Transfer for Long Distance Imaging and Vision Through Scattering Media
Zheng Shi, Ethan Tseng, Mario Bijelic, Werner Ritter, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3476-3486
Adverse weather conditions, including snow, rain, and fog, pose a major challenge for both human and computer vision. Handling these environmental conditions is essential for safe decision making, especially in autonomous vehicles, robotics, and drones. Most of today's supervised imaging and vision approaches, however, rely on training data collected in the real world that is biased towards good weather conditions, with dense fog, snow, and heavy rain as outliers in these datasets. Without training data, let alone paired data, existing autonomous vehicles often limit themselves to good conditions and stop when dense fog or snow is detected. In this work, we tackle the lack of supervised training data by combining synthetic and indirect supervision. We present ZeroScatter, a domain transfer method for converting RGB-only captures taken in adverse weather into clear daytime scenes. ZeroScatter exploits model-based, temporal, multi-view, multi-modal, and adversarial cues in a joint fashion, allowing us to train on unpaired, biased data. We assess the proposed method on in-the-wild captures, and the proposed method outperforms existing monocular descattering approaches by 2.8 dB PSNR on controlled fog chamber measurements.

****
Defending Multimodal Fusion Models Against Single-Source Adversaries
Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, Zico Kolter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3340-3349
Beyond achieving high performance across many vision tasks, multimodal models are expected to be robust to single-source faults due to the availability of redundant information between modalities. In this paper, we investigate the robustness of multimodal neural networks against worst-case (i.e., adversarial) perturbations on a single modality. We first show that standard multimodal fusion models are vulnerable to single-source adversaries: an attack on any single modality can overcome the correct information from multiple unperturbed modalities and cause the model to fail. This surprising vulnerability holds across diverse multimodal tasks and necessitates a solution. Motivated by this finding, we propose an adversarially robust fusion strategy that trains the model to compare information coming from all the input sources, detect inconsistencies in the perturbed modality compared to the other modalities, and only allow information from the unperturbed modalities to pass through. Our approach significantly improves on state-of-the-art methods in single-source robustness, achieving gains of 7.8-25.2% on action recognition, 19.7-48.2% on object detection, and 1.6-6.7% on sentiment analysis, without degrading performance on unperturbed (i.e., clean) data.
****
Generalized Domain Adaptation
Yu Mitsuzumi, Go Irie, Daiki Ikami, Takashi Shibata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1084-1093
Many variants of unsupervised domain adaptation (UDA) problems have been proposed and solved individually. Its side effect is that a method that works for one variant is often ineffective for or not even applicable to another, which has prevented practical applications. In this paper, we give a general representation of UDA problems, named Generalized Domain Adaptation (GDA). GDA covers the major variants as special cases, which allows us to organize them in a comprehensive framework. Moreover, this generalization leads to a new challenging setting where existing methods fail, such as when domain labels are unknown, and class labels are only partially given to each domain. We propose a novel approach to the new setting. The key to our approach is self-supervised class-destructive learning, which enables the learning of class-invariant representations and domain-adversarial classifiers without using any domain labels. Extensive experiments using three benchmark datasets demonstrate that our method outperforms the state-of-the-art UDA methods in the new setting and that it is competitive in existing UDA variations as well.
****
AGORA: Avatars in Geography Optimized for Regression Analysis
Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13468-13478
While the accuracy of 3D human pose estimation from images has steadily improved on benchmark datasets, the best methods still fail in many real-world scenarios. This suggests that there is a domain gap between current datasets and common scenes containing people. To obtain ground-truth 3D pose, current datasets limit the complexity of clothing, environmental conditions, number of subjects, and occlusion. Moreover, current datasets evaluate sparse 3D joint locations corresponding to the major joints of the body, ignoring the hand pose and the face shape. To evaluate the current state-of-the-art methods on more challenging images, and to drive the field to address new problems, we introduce AGORA, a synthetic dataset with high realism and highly accurate ground truth. Here we use 4240 commercially-available, high-quality, textured human scans in diverse poses and natural clothing; this includes 257 scans of children. We create reference 3D poses and body shapes by fitting the SMPL-X body model (with face and hands) to the 3D scans, taking into account clothing. We create around 14K training and 3K test i

mages by rendering between 5 and 15 people per image using either image-based lighting or rendered 3D environments, taking care to make the images physically plausible and photoreal. In total, AGORA consists of 173K individual person crops. We evaluate existing state-of-the-art methods for 3D human pose estimation on this dataset and find that most methods perform poorly on images of children. Hence, we extend the SMPL-X model to better capture the shape of children. Additionally, we fine-tune methods on AGORA and show improved performance on both AGORA and 3DPW, confirming the realism of the dataset. We provide all the registered 3D reference training data, rendered images, and a web-based evaluation site at https://agora.is.tue.mpg.de/.
****
Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation
Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13753-13762
Automatically generating radiology reports can improve current clinical practice in diagnostic radiology. On one hand, it can relieve radiologists from the heavy burden of report writing; On the other hand, it can remind radiologists of abnormalities and avoid the misdiagnosis and missed diagnosis. Yet, this task remains a challenging job for data-driven neural networks, due to the serious visual and textual data biases. To this end, we propose a Posterior-and-Prior Knowledge Exploring-and-Distilling approach (PPKED) to imitate the working patterns of radiologists, who will first examine the abnormal regions and assign the disease topic tags to the abnormal regions, and then rely on the years of prior medical knowledge and prior working experience accumulations to write reports. Thus, the PPKED includes three modules: Posterior Knowledge Explorer (PoKE), Prior Knowledge Explorer (PrKE) and Multi-domain Knowledge Distiller (MKD). In detail, PoKE explores the posterior knowledge, which provides explicit abnormal visual regions to alleviate visual data bias; PrKE explores the prior knowledge from the prior medical knowledge graph (medical knowledge) and prior radiology reports (working experience) to alleviate textual data bias. The explored knowledge is distilled by the MKD to generate the final reports. Evaluated on MIMIC-CXR and IU-Xray datasets, our method is able to outperform previous state-of-the-art models on these two datasets.
****
Rotation Coordinate Descent for Fast Globally Optimal Rotation Averaging
Alvaro Parra, Shin-Fang Chng, Tat-Jun Chin, Anders Eriksson, Ian Reid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4298-4307
Under mild conditions on the noise level of the measurements, rotation averaging satisfies strong duality, which enables global solutions to be obtained via semidefinite programming (SDP) relaxation. However, generic solvers for SDP are rather slow in practice, even on rotation averaging instances of moderate size, thus developing specialised algorithms is vital. In this paper, we present a fast algorithm that achieves global optimality called rotation coordinate descent (RCD). Unlike block coordinate descent (BCD) which solves SDP by updating the semidefinite matrix in a row-by-row fashion, RCD directly maintains and updates all valid rotations throughout the iterations. This obviates the need to store a large dense semidefinite matrix. We mathematically prove the convergence of our algorithm and empirically show its superior efficiency over state-of-the-art global methods on a variety of problem configurations. Maintaining valid rotations also facilitates incorporating local optimisation routines for further speed-ups. Moreover, our algorithm is simple to implement.
****
Extreme Rotation Estimation Using Dense Correlation Volumes
Ruojin Cai, Bharath Hariharan, Noah Snavely, Hadar Averbuch-Elor; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14566-14575
We present a technique for estimating the relative 3D rotation of an RGB image p

air in an extreme setting, where the images have little or no overlap. We observe that, even when images do not overlap, there may be rich hidden cues as to their geometric relationship, such as light source directions, vanishing points, and symmetries present in the scene. We propose a network design that can automatically learn such implicit cues by comparing all pairs of points between the two input images. Our method therefore constructs dense feature correlation volumes and processes these to predict relative 3D rotations. Our predictions are formed over a fine-grained discretization of rotations, bypassing difficulties associated with regressing 3D rotations. We demonstrate our approach on a large variety of extreme RGB image pairs, including indoor and outdoor images captured under different lighting conditions and geographic locations. Our evaluation shows that our model can successfully estimate relative rotations among non-overlapping images without compromising performance over overlapping image pairs.
****
Capsule Network Is Not More Robust Than Convolutional Network
Jindong Gu, Volker Tresp, Han Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14309-14317
The Capsule Network is widely believed to be more robust than Convolutional Networks. However, there lack comprehensive comparisons between these two networks, and it is also unknown which components in the CapsNet affect its robustness. In this paper, we first carefully examine the special designs in CapsNet differing from that of a ConvNet, commonly used for image classification. The examination reveals 5 major new/different components in CapsNet: a transformation process, a dynamic routing layer, a squashing function, a marginal loss other than cross-entropy loss, and an additional class-conditional reconstruction loss for regularization. Along with these major differences, we comprehensively ablate their behavior on 3 kinds of robustness, including affine transformation, overlapping digits, and semantic representation. The study reveals that some designs which are thought critical to CapsNet actually can harm its robustness, i.e., the dynamic routing layer and the transformation process, while others are beneficial for the robustness. Based on these findings, we propose enhanced ConvNets simply by introducing the essential components behind the CapsNet's success. The proposed simple ConvNets can achieve better robustness than the CapsNet.
****
BASAR:Black-Box Attack on Skeletal Action Recognition
Yunfeng Diao, Tianjia Shao, Yong-Liang Yang, Kun Zhou, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7597-7607
Skeletal motion plays a vital role in human activity recognition as either an independent data source or a complement. The robustness of skeleton-based activity recognizers has been questioned recently, which shows that they are vulnerable to adversarial attacks when the full-knowledge of the recognizer is accessible to the attacker. However, this white-box requirement is overly restrictive in most scenarios and the attack is not truly threatening. In this paper, we show that such threats do exist under black-box settings too. To this end, we propose the first black-box adversarial attack method BASAR. Through BASAR, we show that adversarial attack is not only truly a threat but also can be extremely deceitful, because on-manifold adversarial samples are rather common in skeletal motions, in contrast to the common belief that adversarial samples only exist off-manifold. Through exhaustive evaluation and comparison, we show that BASAR can deliver successful attacks across models, data, and attack modes. Through harsh perceptual studies, we show that it achieves effective yet imperceptible attacks. By analyzing the attack on different activity recognizers, BASAR helps identify the potential causes of their vulnerability and provides insights on what classifiers are likely to be more robust against attack.
****
Self-Supervised Learning on 3D Point Clouds by Learning Discrete Generative Models
Benjamin Eckart, Wentao Yuan, Chao Liu, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8248-8257

While recent pre-training tasks on 2D images have proven very successful for transfer learning, pre-training for 3D data remains challenging. In this work, we introduce a general method for 3D self-supervised representation learning that 1) remains agnostic to the underlying neural network architecture, and 2) specifically leverages the geometric nature of 3D point cloud data. The proposed task softly segments 3D points into a discrete number of geometric partitions. A self-supervised loss is formed under the interpretation that these soft partitions implicitly parameterize a latent Gaussian Mixture Model (GMM), and that this generative model establishes a data likelihood function. Our pretext task can therefore be viewed in terms of an encoder-decoder paradigm that squeezes learned representations through an implicitly defined parametric discrete generative model bottleneck. We show that any existing neural network architecture designed for supervised point cloud segmentation can be repurposed for the proposed unsupervised pretext task. By maximizing data likelihood with respect to the soft partitions formed by the unsupervised point-wise segmentation network, learned representations are encouraged to contain compositionally rich geometric information. In tests, we show that our method naturally induces semantic separation in feature space, resulting in state-of-the-art performance on downstream applications like model classification and semantic segmentation.
****
Iso-Points: Optimizing Neural Implicit Surfaces With Hybrid Representations
Wang Yifan, Shihao Wu, Cengiz Oztireli, Olga Sorkine-Hornung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 374-383
Neural implicit functions have emerged as a powerful representation for surfaces in 3D. Such a function can encode a high quality surface with intricate details into the parameters of a deep neural network. However, optimizing for the parameters for accurate and robust reconstructions remains a challenge especially when the input data is noisy or incomplete. In this work, we develop a hybrid neural surface representation that allows us to impose geometry-aware sampling and regularization, which significantly improves the fidelity of reconstructions. We propose to use iso-points as an explicit representation for a neural implicit function. These points are computed and updated on-the-fly during training to capture important geometric features and impose geometric constraints on the optimization. We demonstrate that our method can be adopted to improve state-of-the-art techniques for reconstructing neural implicit surfaces from multi-view images or point clouds. Quantitative and qualitative evaluations show that, compared with existing sampling and optimization methods, our approach allows faster convergence, better generalization, and accurate recovery of details and topology.
****
Dense Relation Distillation With Context-Aware Aggregation for Few-Shot Object Detection
Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10185-10194
Conventional deep learning based methods for object detection require a large amount of bounding box annotations for training, which is expensive to obtain such high quality annotated data. Few-shot object detection, which learns to adapt to novel classes with only a few annotated examples, is very challenging since the fine-grained feature of novel object can be easily overlooked with only a few data available. In this work, aiming to fully exploit features of annotated novel object and capture fine-grained features of query object, we propose Dense Relation Distillation with Context-aware Aggregation (DCNet) to tackle the few-shot detection problem. Built on the meta-learning based framework, Dense Relation Distillation module targets at fully exploiting support features, where support features and query feature are densely matched, covering all spatial locations in a feed-forward fashion. The abundant usage of the guidance information endows model the capability to handle common challenges such as appearance changes and occlusions. Moreover, to better capture scale-aware features, Context-aware Aggregation module adaptively harnesses features from different scales for a more com

prehensive feature representation. Extensive experiments illustrate that our proposed approach achieves state-of-the-art results on PASCAL VOC and MS COCO datasets. Code will be made available at https://github.com/hzhupku/DCNet.
****
End-to-End Human Object Interaction Detection With HOI Transformer
Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11825-11834
We propose HOI Transformer to tackle human object interaction (HOI) detection in an end-to-end manner. Current approaches either decouple HOI task into separated stages of object detection and interaction classification or introduce surrogate interaction problem. In contrast, our method, named HOI Transformer, streamlines the HOI pipeline by eliminating the need for many hand-designed components. HOI Transformer reasons about the relations of objects and humans from global image context and directly predicts HOI instances in parallel. A quintuple matching loss is introduced to force HOI predictions in a unified way. Our method is conceptually much simpler and demonstrates improved accuracy. Without bells and whistles, HOI Transformer achieve 26.61% AP on HICO-DET and 52.9% AProle on V-COCO, surpassing previous methods with the advantage of being much simpler. We hope our approach will serve as a simple and effective alternative for HOI tasks. Code is available at https://github.com/bbepoch/HoiTransformer.
****
How Does Topology Influence Gradient Propagation and Model Performance of Deep Networks With DenseNet-Type Skip Connections?
Kartikeya Bhardwaj, Guihong Li, Radu Marculescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13498-13507
DenseNets introduce concatenation-type skip connections that achieve state-of-the-art accuracy in several computer vision tasks. In this paper, we reveal that the topology of the concatenation-type skip connections is closely related to the gradient propagation which, in turn, enables a predictable behavior of DNNs' test performance. To this end, we introduce a new metric called NN-Mass to quantify how effectively information flows through DNNs. Moreover, we empirically show that NN-Mass also works for other types of skip connections, e.g., for ResNets, Wide-ResNets (WRNs), and MobileNets, which contain addition-type skip connections (i.e., residuals or inverted residuals). As such, for both DenseNet-like CNNs and ResNets/WRNs/MobileNets, our theoretically grounded NN-Mass can identify models with similar accuracy, despite having significantly different size/compute requirements. Detailed experiments on both synthetic and real datasets (e.g., MNIST, CIFAR-10, CIFAR-100, ImageNet) provide extensive evidence for our insights. Finally, the closed-form equation of our NN-Mass enables us to design significantly compressed DenseNets (for CIFAR-10) and MobileNets (for ImageNet) directly at initialization without time-consuming training and/or searching.
****
Multi-Shot Temporal Event Localization: A Benchmark
Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, Philip H. S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12596-12606
Current developments in temporal event or action localization usually target actions captured by a single camera. However, extensive events or actions in the wild may be captured as a sequence of shots by multiple cameras at different positions. In this paper, we propose a new and challenging task called multi-shot temporal event localization, and accordingly, collect a large-scale dataset called MUlti-Shot EventS (MUSES). MUSES has 31,477 event instances for a total of 716 video hours. The core nature of MUSES is the frequent shot cuts, for an average of 19 shots per instance and 176 shots per video, which induces large intra-instance variations. Our comprehensive evaluations show that the state-of-the-art method in temporal action localization only achieves an mAP of 13.1% at IoU=0.5. As a minor contribution, we present a simple baseline approach for handling the intra-instance variations, which reports an mAP of 18.9% on MUSES and 56.9% on THUMOS14 at IoU=0.5. To facilitate research in this direction, we release the datas

et and the project code at https://songbai.site/muses/.
****
We Are More Than Our Joints: Predicting How 3D Bodies Move
Yan Zhang, Michael J. Black, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3372-3382
A key step towards understanding human behavior is the prediction of 3D human motion. Successful solutions have many applications in human tracking, HCI, and graphics. Most previous work focuses on predicting a time series of future 3D joint locations given a sequence 3D joints from the past. This Euclidean formulation generally works better than predicting pose in terms of joint rotations. Body joint locations, however, do not fully constrain 3D human pose, leaving degrees of freedom (like rotation about a limb) undefined. Note that 3D joints can be viewed as a sparse point cloud. Thus the problem of human motion prediction can be seen as a problem of point cloud prediction. With this observation, we instead predict a sparse set of locations on the body surface that correspond to motion capture markers. Given such markers, we fit a parametric body model to recover the 3D body of the person. These sparse surface markers also carry detailed information about human movement that is not present in the joints, increasing the naturalness of the predicted motions. Using the AMASS dataset, we train MOJO (More than Our JOints), which is a novel variational autoencoder with a latent DCT space that generates motions from latent frequencies. MOJO preserves the full temporal resolution of the input motion, and sampling from the latent frequencies explicitly introduces high-frequency components into the generated motion. We note that motion prediction methods accumulate errors over time, resulting in joints or markers that diverge from true human bodies. To address this, we fit the SMPL-X body model to the predictions at each time step, projecting the solution back onto the space of valid bodies, before propagating the new markers in time. Quantitative and qualitative experiments show that our approach produces state-of-the-art results and realistic 3D body animations. The code is available for research purposes at https://yz-cnsdqz.github.io/MOJO/MOJO.html .
****
Spatially-Adaptive Pixelwise Networks for Fast Image Translation
Tamar Rott Shaham, Michael Gharbi, Richard Zhang, Eli Shechtman, Tomer Michaeli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14882-14891
We introduce a new generator architecture, aimed at fast and efficient high-resolution image-to-image translation. We design the generator to be an extremely lightweight function of the full-resolution image. In fact, we use pixel-wise networks; that is, each pixel is processed independently of others, through a composition of simple affine transformations and nonlinearities. We take three important steps to equip such a seemingly simple function with adequate expressivity. First, the parameters of the pixel-wise networks are spatially varying so they can represent a broader function class than simple 1x1 convolutions. Second, these parameters are predicted by a fast convolutional network that processes an aggressively low-resolution representation of the input. Third, we augment the input image with a sinusoidal encoding of spatial coordinates, which provides an effective inductive bias for generating realistic novel high-frequency image content. As a result, our model is up to 18x faster than state-of-the-art baselines. We achieve this speedup while generating comparable visual quality across different image resolutions and translation domains.
****
PointFlow: Flowing Semantics Through Points for Aerial Image Segmentation
Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, Zhouchen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4217-4226
Aerial Image Segmentation is a particular semantic segmentation problem and has several challenging characteristics that general semantic segmentation does not have. There are two critical issues: The one is an extremely foreground-background imbalanced distribution and the other is multiple small objects along with complex background. Such problems make the recent dense affinity context modeling

perform poorly even compared with baselines due to over-introduced background context. To handle these problems, we propose a point-wise affinity propagation module based on the FPN framework, named PointFlow. Rather than dense affinity learning, a sparse affinity map is generated upon selected points between the adjacent features, which reduces the noise introduced by the background while keeping efficiency. In particular, we design a dual point matcher to select points from the salient area and object boundaries, respectively. The former samples salient points while the latter samples points from the object boundaries. Experimental results on three different aerial segmentation datasets suggest that the proposed method is more effective and efficient than state-of-the-art general semantic segmentation methods. Especially, our methods achieve the best speed and accuracy trade-off on three aerial benchmarks. Further experiments on three general semantic segmentation datasets prove the generality of our method. Both code and models will be available for further research.
****
Deep Stable Learning for Out-of-Distribution Generalization
Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5372-5382
Approaches based on deep neural networks have achieved striking performance when testing data and training data share similar distribution, but can significantly fail otherwise. Therefore, eliminating the impact of distribution shifts between training and testing data is crucial for building performance-promising deep models. Conventional methods assume either the known heterogeneity of training data (e.g. domain labels) or the approximately equal capacities of different domains. In this paper, we consider a more challenging case where neither of the above assumptions holds. We propose to address this problem by removing the dependencies between features via learning weights for training samples, which helps deep models get rid of spurious correlations and, in turn, concentrate more on the true connection between discriminative features and labels. Extensive experiments clearly demonstrate the effectiveness of our method on multiple distribution generalization benchmarks compared with state-of-the-art counterparts. Through extensive experiments on distribution generalization benchmarks including PACS, VLCS, MNIST-M, and NICO, we show the effectiveness of our method compared with state-of-the-art counterparts.
****
Continual Learning via Bit-Level Information Preserving
Yujun Shi, Li Yuan, Yunpeng Chen, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16674-16683
Continual learning tackles the setting of learning different tasks sequentially. Despite the lots of previous solutions, most of them still suffer significant forgetting or expensive memory cost. In this work, targeted at these problems, we first study the continual learning process through the lens of information theory and observe that forgetting of a model stems from the loss of information gain on its parameters from the previous tasks when learning a new task. From this viewpoint, we then propose a novel continual learning approach called Bit-Level Information Preserving (BLIP) that preserves the information gain on model parameters through updating the parameters at the bit level, which can be conveniently implemented with parameter quantization. More specifically, BLIP first trains a neural network with weight quantization on the new incoming task and then estimates information gain on each parameter provided by the task data to determine the bits to be frozen to prevent forgetting. We conduct extensive experiments ranging from classification tasks to reinforcement learning tasks, and the results show that our method produces better or on par results comparing to previous state-of-the-arts. Indeed, BLIP achieves close to zero forgetting while only requiring constant memory overheads throughout continual learning.
****
Vectorization and Rasterization: Self-Supervised Learning for Sketch and Handwriting
Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M. Hospedales, T

ao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2021, pp. 5672-5681
Self-supervised learning has gained prominence due to its efficacy at learning p
owerful representations from unlabelled data that achieve excellent performance
on many challenging downstream tasks. However, supervision-free pre-text tasks a
re challenging to design and usually modality specific. Although there is a rich
 literature of self-supervised methods for either spatial (such as images) or te
mporal data (sound or text) modalities, a common pre-text task that benefits bot
h modalities is largely missing. In this paper, we are interested in defining a
self-supervised pre-text task for sketches and handwriting data. This data is un
iquely characterised by its existence in dual modalities of rasterized images an
d vector coordinate sequences. We address and exploit this dual representation b
y proposing two novel cross-modal translation pre-text tasks for self-supervised
 feature learning: Vectorization and Rasterization. Vectorization learns to map
image space to vector coordinates and rasterization maps vector coordinates to i
mage space. We show that our learned encoder modules benefit both raster-based a
nd vector-based downstream approaches to analysing hand-drawn data. Empirical ev
idence shows that our novel pre-text tasks surpass existing single and multi-mod
al self-supervision methods.
****
Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE
Jialun Peng, Dong Liu, Songcen Xu, Houqiang Li; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10775-10784
Given an incomplete image without additional constraint, image inpainting native
ly allows for multiple solutions as long as they appear plausible. Recently, mul
tiple-solution inpainting methods have been proposed and shown the potential of
generating diverse results. However, these methods have difficulty in ensuring t
he quality of each solution, e.g. they produce distorted structure and/or blurry
 texture. We propose a two-stage model for diverse inpainting, where the first s
tage generates multiple coarse results each of which has a different structure,
and the second stage refines each coarse result separately by augmenting texture
. The proposed model is inspired by the hierarchical vector quantized variationa
l auto-encoder (VQ-VAE), whose hierarchical architecture disentangles structural
 and textural information. In addition, the vector quantization in VQ-VAE enable
s autoregressive modeling of the discrete distribution over the structural infor
mation. Sampling from the distribution can easily generate diverse and high-qual
ity structures, making up the first stage of our model. In the second stage, we
propose a structural attention module inside the texture generation network, whe
re the module utilizes the structural information to capture distant correlation
s. We further reuse the VQ-VAE to calculate two feature losses, which help impro
ve structure coherence and texture realism, respectively. Experimental results o
n CelebA-HQ, Places2, and ImageNet datasets show that our method not only enhanc
es the diversity of the inpainting solutions but also improves the visual qualit
y of the generated multiple images. Code and models are available at: https://gi
thub.com/USTC-JialunPeng/Diverse-Structure-Inpainting.
****
Refine Myself by Teaching Myself: Feature Refinement via Self-Knowledge Distilla
tion
Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, Il-Chul Moon; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
021, pp. 10664-10673
Knowledge distillation is a method of transferring the knowledge from a pretrain
ed complex teacher model to a student model, so a smaller network can replace a
large teacher network at the deployment stage. To reduce the necessity of traini
ng a large teacher model, the recent literatures introduced a self-knowledge dis
tillation, which trains a student network progressively to distill its own knowl
edge without a pretrained teacher network. While Self-knowledge distillation is
largely divided into a data augmentation based approach and an auxiliary network
 based approach, the data augmentation approach looses its local information in
the augmentation process, which hinders its applicability to diverse vision task

s, such as semantic segmentation. Moreover, these knowledge distillation approac
hes do not receive the refined feature maps, which are prevalent in the object d
etection and semantic segmentation community. This paper proposes a novel self-k
nowledge distillation method, Feature Refinement via Self-Knowledge Distillation
 (FRSKD), which utilizes an auxiliary self-teacher network to transfer a refined
 knowledge for the classifier network. Our proposed method, FRSKD, can utilize b
oth soft label and feature-map distillations for the self-knowledge distillation
. Therefore, FRSKD can be applied to classification, and semantic segmentation,
which emphasize preserving the local information. We demonstrate the effectivene
ss of FRSKD by enumerating its performance improvements in diverse tasks and ben
chmark datasets. The implemented code will be open-sourced.
****
Self-Supervised Visibility Learning for Novel View Synthesis
Yujiao Shi, Hongdong Li, Xin Yu; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2021, pp. 9675-9684
We address the problem of novel view synthesis (NVS) from a few sparse source vi
ew images. Conventional image-based rendering methods estimate scene geometry an
d synthesize novel views in two separate steps. However, erroneous geometry esti
mation will decrease NVS performance as view synthesis highly depends on the qua
lity of estimated scene geometry. In this paper, we propose an end-to-end NVS fr
amework to eliminate the error propagation issue. To be specific, we construct a
 volume under the target view and design a source-view visibility estimation (SV
E) module to determine the visibility of the target-view voxels in each source v
iew. Next, we aggregate the visibility of all source views to achieve a consensu
s volume. Each voxel in the consensus volume indicates a surface existence proba
bility. Then, we present a soft ray-casting (SRC) mechanism to find the most fro
nt surface in the target view (i.e. depth). Specifically, our SRC traverses the
consensus volume along viewing rays and then estimates a depth probability distr
ibution. We then warp and aggregate source view pixels to synthesize a novel vie
w based on the estimated source-view visibility and target-view depth. At last,
our network is trained in an end-to-end self-supervised fashion, thus significan
tly alleviating error accumulation in view synthesis. Experimental results demon
strate that our method generates novel views in higher quality compared to the s
tate-of-the-art.
****
End-to-End Human Pose and Mesh Reconstruction with Transformers
Kevin Lin, Lijuan Wang, Zicheng Liu; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2021, pp. 1954-1963
We present a new method, called MEsh TRansfOrmer (METRO), to reconstruct 3D huma
n pose and mesh vertices from a single image. Our method uses a transformer enco
der to jointly model vertex-vertex and vertex-joint interactions, and outputs 3D
 joint coordinates and mesh vertices simultaneously. Compared to existing techni
ques that regress pose and shape parameters, METRO does not rely on any parametr
ic mesh models like SMPL, thus it can be easily extended to other objects such a
s hands. We further relax the mesh topology and allow the transformer self-atten
tion mechanism to freely attend between any two vertices, making it possible to
learn non-local relationships among mesh vertices and joints. With the proposed
masked vertex modeling, our method is more robust and effective in handling chal
lenging situations like partial occlusions. METRO generates new state-of-the-art
 results for human mesh reconstruction on the public Human3.6M and 3DPW datasets
. Moreover, we demonstrate the generalizability of METRO to 3D hand reconstructi
on in the wild, outperforming existing state-of-the-art methods on FreiHAND data
set.
****
CapsuleRRT: Relationships-Aware Regression Tracking via Capsules
Ding Ma, Xiangqian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2021, pp. 10948-10957
Regression tracking has gained more and more attention thanks to its easy-to-imp
lement characteristics, while existing regression trackers rarely consider the r
elationships between the object parts and the complete object. This would ultima

tely result in drift from the target object when missing some parts of the target object. Recently, Capsule Network (CapsNet) has shown promising results for image classification benefits from its part-object relationships mechanism, while CapsNet is known for its high computational demand even when carrying out simple tasks. Therefore, a primitive adaptation of CapsNet to regression tracking does not make sense, since this will seriously affect speed of a tracker. To solve these problems, we first explore the spatial-temporal relationships endowed by the CapsNet for regression tracking. The entire regression framework, dubbed CapsuleRRT, consists of three parts. One is S-Caps, which captures the spatial relationships between the parts and the object. Meanwhile, a T-Caps module is designed to exploit the temporal relationships within the target. The response of the target is obtained by STCaps Learning. Further, a prior-guided capsule routing algorithm is proposed to generate more accurate capsule assignments for subsequent frames. Apart from this, the heavy computation burden in CapsNet is addressed with a knowledge distillation pose matrix compression strategy that exploits more tight and discriminative representation with few samples. Extensive experimental results show that CapsuleRRT performs favorably against state-of-the-art methods in terms of accuracy and speed.
****

Test-Time Fast Adaptation for Dynamic Scene Deblurring via Meta-Auxiliary Learning

Zhixiang Chi, Yang Wang, Yuanhao Yu, Jin Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9137-9146

In this paper, we tackle the problem of dynamic scene deblurring. Most existing deep end-to-end learning approaches adopt the same generic model for all unseen test images. These solutions are sub-optimal, as they fail to utilize the internal information within a specific image. On the other hand, a self-supervised approach, SelfDeblur, enables internal-training within a test image from scratch, but it does not fully take advantages of large external dataset. In this work, we propose a novel self-supervised meta-auxiliary learning to improve the performance of deblurring by integrating both external and internal learning. Concretely, we build a self-supervised auxiliary reconstruction task which shares a portion of the network with the primary deblurring task. The two tasks are jointly trained on an external dataset. Furthermore, we propose a meta-auxiliary training scheme to further optimize the pre-trained model as a base learner which is applicable for fast adaptation at test time. During training, the performance of both tasks is coupled. Therefore, we are able to exploit the internal information at test time via the auxiliary task to enhance the performance of deblurring. Extensive experimental results across evaluation datasets demonstrate the effectiveness of test-time adaptation of the proposed method.
****

Anycost GANs for Interactive Image Synthesis and Editing

Ji Lin, Richard Zhang, Frieder Ganz, Song Han, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14986-14996

Generative adversarial networks (GANs) have enabled photorealistic image synthesis and editing. However, due to the high computational cost of large-scale generators (e.g., StyleGAN2), it usually takes seconds to see the results of a single edit on edge devices, prohibiting interactive user experience. In this paper, inspired by quick preview features in modern rendering software, we propose Anycost GAN for interactive natural image editing. We train the Anycost GAN to support elastic resolutions and channels for faster image generation at versatile speeds. Running subsets of the full generator produce outputs that are perceptually similar to the full generator, making them a good proxy for a quick preview. By using sampling-based multi-resolution training, adaptive-channel training, and a generator-conditioned discriminator, the anycost generator can be evaluated at various configurations while achieving better image quality compared to separately trained models. Furthermore, we develop new encoder training and latent code optimization techniques to encourage consistency between the different sub-generators during image projection. Anycost GAN can be executed at various cost budge

ts (up to 10x computation reduction) and adapt to a wide range of hardware and l atency requirements. When deployed on desktop CPUs and edge devices, our model c an provide perceptually similar previews at 6-12x speedup, enabling interactive image editing. The code and demo are publicly available.
****
TrafficSim: Learning To Simulate Realistic Multi-Agent Behaviors
Simon Suo, Sebastian Regalado, Sergio Casas, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10400-10409
Simulation has the potential to massively scale evaluation of self-driving syste ms, enabling rapid development as well as safe deployment. Bridging the gap betw een simulation and the real world requires realistic multi-agent behaviors. Exis ting simulation environments rely on heuristic-based models that directly encode traffic rules, which cannot capture irregular maneuvers (e.g., nudging, U-turns) and complex interactions (e.g., yielding, merging). In contrast, we leverage r eal-world data to learn directly from human demonstration, and thus capture more naturalistic driving behaviors. To this end, we propose TrafficSim, a multi-age nt behavior model for realistic traffic simulation. In particular, we parameteri ze the policy with an implicit latent variable model that generates socially-con sistent plans for all actors in the scene jointly. To learn a robust policy amen able for long horizon simulation, we unroll the policy in training and optimize through the fully differentiable simulation across time. Our learning objective incorporates both human demonstrations as well as common sense. We show TrafficS im generates significantly more realistic traffic scenarios as compared to a div erse set of baselines. Notably, we can exploit trajectories generated by Traffic Sim as effective data augmentation for training better motion planner.
****
Monocular 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks
Yu Cheng, Bo Wang, Bo Yang, Robby T. Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7649-7659
In monocular video 3D multi-person pose estimation, inter-person occlusion and c lose interactions can cause human detection to be erroneous and human-joints gro uping to be unreliable. Existing top-down methods rely on human detection and th us suffer from these problems. Existing bottom-up methods do not use human detec tion, but they process all persons at once at the same scale, causing them to be sensitive to multiple-persons scale variations. To address these challenges, we propose the integration of top-down and bottom-up approaches to exploit their s trengths. Our top-down network estimates human joints from all persons instead o f one in an image patch, making it robust to possible erroneous bounding boxes. Our bottom-up network incorporates human-detection based normalized heatmaps, al lowing the network to be more robust in handling scale variations. Finally, the estimated 3D poses from the top-down and bottom-up networks are fed into our int egration network for final 3D poses. Besides the integration of top-down and bot tom-up networks, unlike existing pose discriminators that are designed solely fo r single person, and consequently cannot assess natural inter-person interaction s, we propose a two-person pose discriminator that enforces natural two-person i nteractions. Lastly, we also apply a semi-supervised method to overcome the 3D g round-truth data scarcity. Our quantitative and qualitative evaluations show the effectiveness of our method compared to the state-of-the-art baselines.
****
Space-Time Distillation for Video Super-Resolution
Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2113-2122
Compact video super-resolution (VSR) networks can be easily deployed on resource -limited devices, e.g., smart-phones and wearable devices, but have considerable performance gaps compared with complicated VSR networks that require a large am ount of computing resources. In this paper, we aim to improve the performance of compact VSR networks without changing their original architectures, through a k

nowledge distillation approach that transfers knowledge from a complicated VSR n
etwork to a compact one. Specifically, we propose a space-time distillation (STD
) scheme to exploit both spatial and temporal knowledge in the VSR task. For spa
ce distillation, we extract spatial attention maps that hints the high-frequency
 video content from both networks, which are further used for transferring spati
al modeling ability. For time distillation, we narrow the performance gap betwee
n compact models and complicated models by distilling the feature similarity of
the temporal memory cells, which is encoded from the sequence of feature maps ge
nerated in the training clips using ConvLSTM. During the training process, STD c
an be easily incorporated into any network without changing the original network
 architecture. Experimental results on standard benchmarks demonstrate that, in
resource-constrained situations, the proposed method notably improve the perform
ance of existing VSR networks without increasing the inference time.
****
Robust Audio-Visual Instance Discrimination
Pedro Morgado, Ishan Misra, Nuno Vasconcelos; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12934-12945
We present a self-supervised learning method to learn audio and video representa
tions. Prior work uses the natural correspondence between audio and video to def
ine a standard cross-modal instance discrimination task, where a model is traine
d to match representations from the two modalities. However, the standard approa
ch introduces two sources of training noise. First, audio-visual correspondences
 often produce faulty positives since the audio and video signals can be uninfor
mative of each other. To limit the detrimental impact of faulty positives, we op
timize a weighted contrastive learning loss, which down-weighs their contributio
n to the overall loss. Second, since self-supervised contrastive learning relies
 on random sampling of negative instances, instances that are semantically simil
ar to the base instance can be used as faulty negatives. To alleviate the impact
 of faulty negatives, we propose to optimize an instance discrimination loss wit
h a soft target distribution that estimates relationships between instances. We
validate our contributions through extensive experiments on action recognition t
asks and show that they address the problems of audio-visual instance discrimina
tion and improve transfer learning performance.
****
High-Fidelity and Arbitrary Face Editing
Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, Zhouhui Lian
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2021, pp. 16115-16124
Cycle consistency is widely used for face editing. However, we observe that the
generator tends to find a tricky way to hide information from the original image
 to satisfy the constraint of cycle consistency, making it impossible to maintai
n the rich details (e.g., wrinkles and moles) of nonediting areas. In this work,
 we propose a simple yet effective method named HifaFace to address the above-me
ntioned problem from two perspectives. First, we relieve the pressure of the gen
erator to synthesize rich details by directly feeding the high-frequency informa
tion of the input image into the end of the generator. Second, we adopt an addit
ional discriminator to encourage the generator to synthesize rich details. Speci
fically, we apply wavelet transformation to transform the image into multi-frequ
ency domains, among which the high-frequency parts can be used to recover the ri
ch details. We also notice that a fine-grained and wider-range control for the a
ttribute is of great importance for face editing. To achieve this goal, we propo
se a novel attribute regression loss. Powered by the proposed framework, we achi
eve high-fidelity and arbitrary face editing, outperforming other state-of-the-a
rt approaches.
****
Explicit Knowledge Incorporation for Visual Reasoning
Yifeng Zhang, Ming Jiang, Qi Zhao; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2021, pp. 1356-1365
Existing explainable and explicit visual reasoning methods only perform reasonin
g based on visual evidence but do not take into account knowledge beyond what is

in the visual scene. To addresses the knowledge gap between visual reasoning methods and the semantic complexity of real-world images, we present the first explicit visual reasoning method that incorporates external knowledge and models high-order relational attention for improved generalizability and explainability. Specifically, we propose a knowledge incorporation network that explicitly creates and includes new graph nodes for entities and predicates from external knowledge bases to enrich the semantics of the scene graph used in explicit reasoning. We then create a novel Graph-Relate module to perform high-order relational attention on the enriched scene graph. By explicitly introducing structured external knowledge and high-order relational attention, our method demonstrates significant generalizability and explainability over the state-of-the-art visual reasoning approaches on the GQA and VQAv2 datasets.
****
Progressive Unsupervised Learning for Visual Object Tracking
Qiangqiang Wu, Jia Wan, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2993-3002
In this paper, we propose a progressive unsupervised learning (PUL) framework, which entirely removes the need for annotated training videos in visual tracking. Specifically, we first learn a background discrimination (BD) model that effectively distinguishes an object from background in a contrastive learning way. We then employ the BD model to progressively mine temporal corresponding patches (i.e., patches connected by a track) in sequential frames. As the BD model is imperfect and thus the mined patch pairs are noisy, we propose a noise-robust loss function to more effectively learn temporal correspondences from this noisy data. We use the proposed noise robust loss to train backbone networks of Siamese trackers. Without online fine-tuning or adaptation, our unsupervised real-time Siamese trackers can outperform state-of-the-art unsupervised deep trackers and achieve competitive results to the supervised baselines.
****
IoU Attack: Towards Temporally Coherent Black-Box Adversarial Attack for Visual Object Tracking
Shuai Jia, Yibing Song, Chao Ma, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6709-6718
Adversarial attack arises due to the vulnerability of deep neural networks to perceive input samples injected with imperceptible perturbations. Recently, adversarial attack has been applied to visual object tracking to evaluate the robustness of deep trackers. Assuming that the model structures of deep trackers are known, a variety of white-box attack approaches to visual tracking have demonstrated promising results. However, the model knowledge about deep trackers is usually unavailable in real applications. In this paper, we propose a decision-based black-box attack method for visual object tracking. In contrast to existing black-box adversarial attack methods that deal with static images for image classification, we propose IoU attack that sequentially generates perturbations based on the predicted IoU scores from both current and historical frames. By decreasing the IoU scores, the proposed attack method degrades the accuracy of temporal coherent bounding boxes (i.e., object motions) accordingly. In addition, we transfer the learned perturbations to the next few frames to initialize temporal motion attacks. We validate the proposed IoU attack on state-of-the-art deep trackers (i.e., detection based, correlation filter based, and long-term trackers). Extensive experiments on the benchmark datasets indicate the effectiveness of the proposed IoU attack method. The source code is available at https://github.com/VISION-SJTU/IoUattack.
****
Deep Graph Matching Under Quadratic Constraint
Quankai Gao, Fudong Wang, Nan Xue, Jin-Gang Yu, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5069-5078
Recently, deep learning based methods have demonstrated promising results on the graph matching problem, by relying on the descriptive capability of deep features extracted on graph nodes. However, one main limitation with existing deep gra

ph matching (DGM) methods lies in their ignorance of explicit constraint of grap
h structures, which may lead the model to be trapped into local minimum in train
ing. In this paper, we propose to explicitly formulate pairwise graph structures
 as a quadratic constraint incorporated into the DGM framework. The quadratic co
nstraint minimizes the pairwise structural discrepancy between graphs, which can
 reduce the ambiguities brought by only using the extracted CNN features. Moreov
er, we present a differentiable implementation to the quadratic constrained-opti
mization such that it is compatible with the unconstrained deep learning optimiz
er. To give more precise and proper supervision, a well-designed false matching
loss against class imbalance is proposed, which can better penalize the false ne
gatives and false positives with less overfitting. Exhaustive experiments demons
trate that our method achieves competitive performance on real-world datasets. T
he code is available at: https://github.com/Zerg-Overmind/QC-DGM.
****
Multi-Label Activity Recognition Using Activity-Specific Features and Activity C
orrelations
Yanyi Zhang, Xinyu Li, Ivan Marsic; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2021, pp. 14625-14635
Multi-label activity recognition is designed for recognizing multiple activities
 that are performed simultaneously or sequentially in each video. Most recent ac
tivity recognition networks focus on single-activities, that assume only one act
ivity in each video. These networks extract shared features for all the activiti
es, which are not designed for multi-label activities. We introduce an approach
to multi-label activity recognition that extracts independent feature descriptor
s for each activity and learns activity correlations. This structure can be trai
ned end-to-end and plugged into any existing network structures for video classi
fication. Our method outperformed state-of-the-art approaches on four multi-labe
l activity recognition datasets. To better understand the activity-specific feat
ures that the system generated, we visualized these activity-specific features i
n the Charades dataset. The code will be released later.
****
Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance V
ideos
Yasamin Jafarian, Hyun Soo Park; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2021, pp. 12753-12762
A key challenge of learning the geometry of dressed humans lies in the limited a
vailability of the ground truth data (e.g., 3D scanned models), which results in
 the performance degradation of 3D human reconstruction when applying to real wo
rld imagery. We address this challenge by leveraging a new data resource: a numb
er of social media dance videos that span diverse appearance, clothing styles, p
erformances, and identities. Each video depicts dynamic movements of the body an
d clothes of a single person while lacking the 3D ground truth geometry. To util
ize these videos, we present a new method to use the local transformation that w
arps the predicted local geometry of the person from an image to that of the oth
er image at a different time instant. With the transformation, the predicted geo
metry can be self-supervised by the warped geometry from the other image. In add
ition, we jointly learn the depth along with the surface normals, which are high
ly responsive to local texture, wrinkle, and shade by maximizing their geometric
 consistency. Our method is end-to-end trainable, resulting in high fidelity dep
th estimation that predicts fine geometry faithful to the input real image. We d
emonstrate that our method outperforms the state-of-the-art human depth estimati
on and human shape recovery approaches on both real and rendered images.
****
Unpaired Image-to-Image Translation via Latent Energy Transport
Yang Zhao, Changyou Chen; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2021, pp. 16418-16427
Image-to-image translation aims to preserve source contents while translating to
 discriminative target styles between two visual domains. Most works apply adver
sarial learning in the ambient image space, which could be computationally expen
sive and challenging to train. In this paper, we propose to deploy an energy-bas

ed model (EBM) in the latent space of a pretrained autoencoder for this task. The pretrained autoencoder serves as both a latent code extractor and an image reconstruction worker. Our model, LETIT, is based on the assumption that two domains share the same latent space, where latent representation is implicitly decomposed as a content code and a domain-specific style code. Instead of explicitly extracting the two codes and applying adaptive instance normalization to combine them, our latent EBM can implicitly learn to transport the source style code to the target style code while preserving the content code, an advantage over existing image translation methods. This simplified solution is also more efficient in the one-sided unpaired image translation setting. Qualitative and quantitative comparisons demonstrate superior translation quality and faithfulness for content preservation. Our model is the first to be applicable to 1024x1024-resolution unpaired image translation to the best of our knowledge. Code is available at https://github.com/YangNaruto/latent-energy-transport.
****

VLN BERT: A Recurrent Vision-and-Language BERT for Navigation
Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1643-1653
Accuracy of many visiolinguistic tasks has benefited significantly from the application of vision-and-language (V&L) BERT. However, its application for the task of vision-and-language navigation (VLN) remains limited. One reason for this is the difficulty adapting the BERT architecture to the partially observable Markov decision process present in VLN, requiring history-dependent attention and decision making. In this paper we propose a recurrent BERT model that is time-aware for use in VLN. Specifically, we equip the BERT model with a recurrent function that maintains cross-modal state information for the agent. Through extensive experiments on R2R and REVERIE we demonstrate that our model can replace more complex encoder-decoder models to achieve state-of-the-art results. Moreover, our approach can be generalised to other transformer-based architectures, supports pre-training, and is capable of solving navigation and referring expression tasks simultaneously.
****

Content-Aware GAN Compression
Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Federico Perazzi, Sun-Yuan Kung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12156-12166
Generative adversarial networks (GANs), e.g., StyleGAN2, play a vital role in various image generation and synthesis tasks, yet their notoriously high computational cost hinders their efficient deployment on edge devices. Directly applying generic compression approaches yields poor results on GANs, which motivates a number of recent GAN compression works. While prior works mainly accelerate conditional GANs, e.g., pix2pix and CycleGAN, compressing state-of-the-art unconditional GANs has rarely been explored and is more challenging. In this paper, we propose novel approaches for unconditional GAN compression. We first introduce effective channel pruning and knowledge distillation schemes specialized for unconditional GANs. We then propose a novel content-aware method to guide the processes of both pruning and distillation. With content-awareness, we can effectively prune channels that are unimportant to the contents of interest, e.g., human faces, and focus our distillation on these regions, which significantly enhances the distillation quality. On StyleGAN2 and SN-GAN, we achieve a substantial improvement over the state-of-the-art compression method. Notably, we reduce the FLOPs of StyleGAN2 by 11x with visually negligible image quality loss compared to the full-size model. More interestingly, when applied to various image manipulation tasks, our compressed model forms a smoother and better disentangled latent manifold, making it more effective for image editing.
****

FBI-Denoiser: Fast Blind Image Denoiser for Poisson-Gaussian Noise
Jaeseok Byun, Sungmin Cha, Taesup Moon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5768-5777

We consider the challenging blind denoising problem for Poisson-Gaussian noise, in which no additional information about clean images or noise level parameters is available. Particularly, when only "single" noisy images are available for training a denoiser, the denoising performance of existing methods was not satisfactory. Recently, the blind pixelwise affine image denoiser (BP-AIDE) was proposed and significantly improved the performance in the above setting, to the extent that it is competitive with denoisers which utilized additional information. However, BP-AIDE seriously suffered from slow inference time due to the inefficiency of noise level estimation procedure and that of the blind-spot network (BSN) architecture it used. To that end, we propose Fast Blind Image Denoiser (FBI-Denoiser) for Poisson-Gaussian noise, which consists of two neural network models; 1) PGE-Net that estimates Poisson-Gaussian noise parameters 2000 times faster than the conventional methods and 2) FBI-Net that realizes a much more efficient BSN for pixelwise affine denoiser in terms of the number of parameters and inference speed. Consequently, we show that our FBI-Denoiser blindly trained solely based on single noisy images can achieve the state-of-the-art performance on several real-world noisy image benchmark datasets with much faster inference time (X 10), compared to BP-AIDE.
****
Hijack-GAN: Unintended-Use of Pretrained, Black-Box GANs
Hui-Po Wang, Ning Yu, Mario Fritz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7872-7881
While Generative Adversarial Networks (GANs) show increasing performance and the level of realism is becoming indistinguishable from natural images, this also comes with high demands on data and computation. We show that state-of-the-art GAN models -- such as they are being publicly released by researchers and industry -- can be used for a range of applications beyond unconditional image generation. We achieve this by an iterative scheme that also allows gaining control over the image generation process despite the highly non-linear latent spaces of the latest GAN models. We demonstrate that this opens up the possibility to re-use state-of-the-art, difficult to train, pre-trained GANs with a high level of control even if only black-box access is granted. Our work also raises concerns and awareness that the use cases of a published GAN model may well reach beyond the creators' intention, which needs to be taken into account before a full public release. Code is available at https://github.com/a514514772/hijackgan.
****
LiDAR R-CNN: An Efficient and Universal 3D Object Detector
Zhichao Li, Feng Wang, Naiyan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7546-7555
LiDAR-based 3D detection in point cloud is essential in the perception system of autonomous driving. In this paper, we present LiDAR R-CNN, a second stage detector that can generally improve any existing 3D detector. To fulfill the real-time and high precision requirement in practice, we resort to point-based approach other than the popular voxel-based approach. However, we find an overlooked issue in previous work: Naively applying point-based methods like PointNet could make the learned features ignore the size of proposals. To this end, we analyze this problem in detail and propose several methods to remedy it, which bring significant performance improvement. Comprehensive experimental results on real-world datasets like Waymo Open Dataset (WOD) and KITTI dataset with various popular detectors demonstrate the universality and superiority of our LiDAR R-CNN. In particular, based on one variant of PointPillars, our method could achieve new state-of-the-art results with minor cost. Codes will be released at https://github.com/tusimple/LiDAR_RCNN.
****
Line Segment Detection Using Transformers Without Edges
Yifan Xu, Weijian Xu, David Cheung, Zhuowen Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4257-4266
In this paper, we present a joint end-to-end line segment detection algorithm using Transformers that is post-processing and heuristics-guided intermediate processing (edge/junction/region detection) free. Our method, named LinE segment TRa

nsformers (LETR), takes advantages of having integrated tokenized queries, a sel f-attention mechanism, and encoding-decoding strategy within Transformers by ski pping standard heuristic designs for the edge element detection and perceptual g rouping processes. We equip Transformers with a multi-scale encoder/decoder stra tegy to perform fine-grained line segment detection under a direct endpoint dist ance loss. This loss term is particularly suitable for detecting geometric struc tures such as line segments that are not conveniently represented by the standar d bounding box representations. The Transformers learn to gradually refine line segments through layers of self-attention. In our experiments, we show state-of-the-art results on Wireframe and YorkUrban benchmarks.
****
Region-Aware Adaptive Instance Normalization for Image Harmonization
Jun Ling, Han Xue, Li Song, Rong Xie, Xiao Gu; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9361-9370
Image composition plays a common but important role in photo editing. To acquire photo-realistic composite images, one must adjust the appearance and visual sty le of the foreground to be compatible with the background. Existing deep learnin g methods for harmonizing composite images directly learn an image mapping netwo rk from the composite to real one, without explicit exploration on visual style consistency between the background and the foreground images. To ensure the visu al style consistency between the foreground and the background, in this paper, w e treat image harmonization as a style transfer problem. In particular, we propo se a simple yet effective Region-aware Adaptive Instance Normalization (RAIN) mo dule, which explicitly formulates the visual style from the background and adapt ively applies them to the foreground. With our settings, our RAIN module can be used as a drop-in module for existing image harmonization networks and is able t o bring significant improvements. Extensive experiments on the existing image ha rmonization benchmark datasets show the superior capability of the proposed meth od. Code is available at https://github.com/junleen/RainNet .
****
Learning Tensor Low-Rank Prior for Hyperspectral Image Reconstruction
Shipeng Zhang, Lizhi Wang, Lei Zhang, Hua Huang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12006-12015
Snapshot hyperspectral imaging has been developed to capture the spectral inform ation of dynamic scenes. In this paper, we propose a deep neural network by lear ning the tensor low-rank prior of hyperspectral images (HSI) in the feature doma in to promote the reconstruction quality. Our method is inspired by the canonica l-polyadic (CP) decomposition theory, where a low-rank tensor can be expressed a s a weight summation of several rank-1 component tensors. Specifically, we first learn the tensor low-rank prior of the image features with two steps: (a) we ge nerate rank-1 tensors with discriminative components to collect the contextual i nformation from both spatial and channel dimensions of the image features; (b) w e aggregate those rank-1 tensors into a low-rank tensor as a 3D attention map to exploit the global correlation and refine the image features. Then, we integrat e the learned tensor low-rank prior into an iterative optimization algorithm to obtain an end-to-end HSI reconstruction. Experiments on both synthetic and real data demonstrate the superiority of our method.
****
Unsupervised Learning of Depth and Depth-of-Field Effect From Natural Images Wit h Aperture Rendering Generative Adversarial Networks
Takuhiro Kaneko; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2021, pp. 15679-15688
Understanding the 3D world from 2D projected natural images is a fundamental cha llenge in computer vision and graphics. Recently, an unsupervised learning appro ach has garnered considerable attention owing to its advantages in data collecti on. However, to mitigate training limitations, typical methods need to impose as sumptions for viewpoint distribution (e.g., a dataset containing various viewpoi nt images) or object shape (e.g., symmetric objects). These assumptions often re strict applications; for instance, the application to non-rigid objects or image s captured from similar viewpoints (e.g., flower or bird images) remains a chall

enge. To complement these approaches, we propose aperture rendering generative a dversarial networks (AR-GANs), which equip aperture rendering on top of GANs, an d adopt focus cues to learn the depth and depth-of-field (DoF) effect of unlabel ed natural images. To address the ambiguities triggered by unsupervised setting (i.e., ambiguities between smooth texture and out-of-focus blurs, and between fo reground and background blurs), we develop DoF mixture learning, which enables t he generator to learn real image distribution while generating diverse DoF image s. In addition, we devise a center focus prior to guiding the learning direction . In the experiments, we demonstrate the effectiveness of AR-GANs in various dat asets, such as flower, bird, and face images, demonstrate their portability by i ncorporating them into other 3D representation learning GANs, and validate their applicability in shallow DoF rendering.
****
Sign-Agnostic Implicit Learning of Surface Self-Similarities for Shape Modeling and Reconstruction From Raw Point Clouds
Wenbin Zhao, Jiabao Lei, Yuxin Wen, Jianguo Zhang, Kui Jia; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10256-10265
Shape modeling and reconstruction from raw point clouds of objects stand as a fu ndamental challenge in vision and graphics research. Classical methods consider analytic shape priors; however, their performance is degraded when the scanned p oints deviate from the ideal conditions of cleanness and completeness. Important progress has been recently made by data-driven approaches, which learn global a nd/or local models of implicit surface representations from auxiliary sets of tr aining shapes. Motivated from a universal phenomenon that self-similar shape pat terns of local surface patches repeat across the entire surface of an object, we aim to push forward the data-driven strategies and propose to learn a local imp licit surface network for a shared, adaptive modeling of the entire surface for a direct surface reconstruction from raw point cloud; we also enhance the levera ging of surface self-similarities by improving correlations among the optimized latent codes of individual surface patches. Given that orientations of raw point s could be unavailable or noisy, we extend signagnostic learning into our local implicit model, which enables our recovery of signed implicit fields of local su rfaces from the unsigned inputs. We term our framework as Sign-Agnostic Implicit Learning of Surface Self-Similarities (SAIL-S3). With a global post-optimizatio n of local sign flipping, SAIL-S3 is able to directly model raw, un-oriented poi nt clouds and reconstruct high-quality object surfaces. Experiments show its sup eriority over existing methods.
****
Towards More Flexible and Accurate Object Tracking With Natural Language: Algori thms and Benchmark
Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2021, pp. 13763-13773
Tracking by natural language specification is a new rising research topic that a ims at locating the target object in the video sequence based on its language de scription. Compared with traditional bounding box (BBox) based tracking, this se tting guides object tracking with high-level semantic information, addresses the ambiguity of BBox, and links local and global search organically together. Thos e benefits may bring more flexible, robust and accurate tracking performance in practical scenarios. However, existing natural language initialized trackers are developed and compared on benchmark datasets proposed for tracking-by-BBox, whi ch can't reflect the true power of tracking-by-language. In this work, we propos e a new benchmark specifically dedicated to the tracking-by-language, including a large scale dataset, strong and diverse baseline methods. Specifically, we col lect 2k video sequences (contains a total of 1,244,340 frames, 663 words) and sp lit 1300/700 for the train/testing respectively. We densely annotate one sentenc e in English and corresponding bounding boxes of the target object for each vide o. We also introduce two new challenges into TNL2K for the object tracking task, i.e., adversarial samples and modality switch. A strong baseline method based o

n an adaptive local-global-search scheme is proposed for future works to compare. We believe this benchmark will greatly boost related researches on natural language guided tracking.
****

On Learning the Geodesic Path for Incremental Learning

Christian Simon, Piotr Koniusz, Mehrtash Harandi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1591-1600

Neural networks notoriously suffer from the problem of catastrophic forgetting, the phenomenon of forgetting the past knowledge when acquiring new knowledge. Overcoming catastrophic forgetting is of significant importance to emulate the process of "incremental learning", where the model is capable of learning from sequential experience in an efficient and robust way. State-of-the-art techniques for incremental learning make use of knowledge distillation towards preventing catastrophic forgetting. Therein, one updates the network while ensuring that the network's responses to previously seen concepts remain stable throughout updates. This in practice is done by minimizing the dissimilarity between current and previous responses of the network one way or another. Our work contributes a novel method to the arsenal of distillation techniques. In contrast to previous state of the art, we propose to firstly construct low-dimensional manifolds for previous and current responses and minimize the dissimilarity between the responses along the geodesic connecting the manifolds. This induces a more formidable knowledge distillation with smooth properties which preserves the past knowledge more efficiently as observed by our comprehensive empirical study.
****

The Lottery Tickets Hypothesis for Supervised and Self-Supervised Pre-Training in Computer Vision Models

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, Zhangyang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16306-16316

The computer vision world has been re-gaining enthusiasm in various pre-trained models, including both classical ImageNet supervised pre-training and recently emerged self-supervised pre-training such as simCLR and MoCo. Pre-trained weights often boost a wide range of downstream tasks including classification, detection, and segmentation. Latest studies suggest that pre-training benefits from gigantic model capacity. We are hereby curious and ask: after pre-training, does a pre-trained model indeed have to stay large for its downstream transferability? In this paper, we examine supervised and self-supervised pre-trained models through the lens of the lottery ticket hypothesis (LTH). LTH identifies highly sparse matching subnetworks that can be trained in isolation from (nearly) scratch yet still reach the full models' performance. We extend the scope of LTH and question whether matching subnetworks still exist in pre-trained computer vision models, that enjoy the same downstream transfer performance. Our extensive experiments convey an overall positive message: from all pre-trained weights obtained by ImageNet classification, simCLR, and MoCo, we are consistently able to locate such matching subnetworks at 59.04% to 96.48% sparsity that transfer universally to multiple downstream tasks, whose performance see no degradation compared to using full pre-trained weights. Further analyses reveal that subnetworks found from different pre-training tend to yield diverse mask structures and perturbation sensitivities. We conclude that the core LTH observations remain generally relevant in the pre-training paradigm of computer vision, but more delicate discussions are needed in some cases. Codes and pre-trained models will be made available at: https://github.com/VITA-Group/CV_LTH_Pre-training.
****

Iterative Shrinking for Referring Expression Grounding Using Deep Reinforcement Learning

Mingjie Sun, Jimin Xiao, Eng Gee Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14060-14069

In this paper, we are tackling the proposal-free referring expression grounding task, aiming at localizing the target object according to a query sentence, without relying on off-the-shelf object proposals. Existing proposal-free methods em

ploy a query-image matching branch to select the highest-score point in the imag
e feature map as the target box center, with its width and height predicted by a
nother branch. Such methods, however, fail to utilize the contextual relation be
tween the target and reference objects, and lack interpretability on its reasoni
ng procedure. To solve these problems, we propose an iterative shrinking mechani
sm to localize the target, where the shrinking direction is decided by a reinfor
cement learning agent, with all contents within the current image patch comprehe
nsively considered. Beside, the sequential shrinking process enables to demonstr
ate the reasoning about how to iteratively find the target. Experiments show tha
t the proposed method boosts the accuracy by 4.32% against the previous state-of
-the-art (SOTA) method on the RefCOCOg dataset, where query sentences are long a
nd complex, with many targets referred by other reference objects.
****
Simulating Unknown Target Models for Query-Efficient Black-Box Attacks
Chen Ma, Li Chen, Jun-Hai Yong; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 11835-11844
Many adversarial attacks have been proposed to investigate the security issues o
f deep neural networks. In the black-box setting, current model stealing attacks
 train a substitute model to counterfeit the functionality of the target model.
However, the training requires querying the target model. Consequently, the quer
y complexity remains high, and such attacks can be defended easily. This study a
ims to train a generalized substitute model called "Simulator", which can mimic
the functionality of any unknown target model. To this end, we build the trainin
g data with the form of multiple tasks by collecting query sequences generated d
uring the attacks of various existing networks. The learning process uses a mean
 square error-based knowledge-distillation loss in the meta-learning to minimize
 the difference between the Simulator and the sampled networks. The meta-gradien
ts of this loss are then computed and accumulated from multiple tasks to update
the Simulator and subsequently improve generalization. When attacking a target m
odel that is unseen in training, the trained Simulator can accurately simulate i
ts functionality using its limited feedback. As a result, a large fraction of qu
eries can be transferred to the Simulator, thereby reducing query complexity. Re
sults of the comprehensive experiments conducted using the CIFAR-10, CIFAR-100,
and TinyImageNet datasets demonstrate that the proposed approach reduces query c
omplexity by several orders of magnitude compared to the baseline method. The im
plementation source code is released online at https://github.com/machanic/Simul
atorAttack.
****
Diffusion Probabilistic Models for 3D Point Cloud Generation
Shitong Luo, Wei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2021, pp. 2837-2845
We present a probabilistic model for point cloud generation, which is fundamenta
l for various 3D vision tasks such as shape completion, upsampling, synthesis an
d data augmentation. Inspired by the diffusion process in non-equilibrium thermo
dynamics, we view points in point clouds as particles in a thermodynamic system
in contact with a heat bath, which diffuse from the original distribution to a n
oise distribution. Point cloud generation thus amounts to learning the reverse d
iffusion process that transforms the noise distribution to the distribution of a
 desired shape. Specifically, we propose to model the reverse diffusion process
for point clouds as a Markov chain conditioned on certain shape latent. We deriv
e the variational bound in closed form for training and provide implementations
of the model. Experimental results demonstrate that our model achieves competiti
ve performance in point cloud generation and auto-encoding. The code is availabl
e at https://github.com/luost26/diffusion-point-cloud
****
Dual Pixel Exploration: Simultaneous Depth Estimation and Image Restoration
Liyuan Pan, Shah Chowdhury, Richard Hartley, Miaomiao Liu, Hongguang Zhang, Hong
dong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2021, pp. 4340-4349
The dual-pixel (DP) hardware works by splitting each pixel in half and creating

an image pair in a single snapshot. Several works estimate depth/inverse depth by treating the DP pair as a stereo pair. However, dual-pixel disparity only occurs in image regions with the defocus blur. The heavy defocus blur in DP pairs affects the performance of matching-based depth estimation approaches. Instead of removing the blur effect blindly, we study the formation of the DP pair which links the blur and the depth information. In this paper, we propose a mathematical DP model which can benefit depth estimation by the blur. These explorations motivate us to propose an end-to-end DDDNet (DP-based Depth and Deblur Network) to jointly estimate the depth and restore the image. Moreover, we define a reblur loss, which reflects the relationship of the DP image formation process with depth information, to regularise our depth estimate in training. To meet the requirement of a large amount of data for learning, we propose the first DP image simulator which allows us to create datasets with DP pairs from any existing RGBD dataset. As a side contribution, we collect a real dataset for further research. Extensive experimental evaluation on both synthetic and real datasets shows that our approach achieves competitive performance compared to state-of-the-art approaches.
****
Guided Integrated Gradients: An Adaptive Path Method for Removing Noise
Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, Tolga Bolukbasi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5050-5058
Integrated Gradients (IG) is a commonly used feature attribution method for deep neural networks. While IG has many desirable properties, the method often produces spurious/noisy pixel attributions in regions that are not related to the predicted class when applied to visual models. While this has been previously noted, most existing solutions are aimed at addressing the symptoms by explicitly reducing the noise in the resulting attributions. In this work, we show that one of the causes of the problem is the accumulation of noise along the IG path. To minimize the effect of this source of noise, we propose adapting the attribution path itself -- conditioning the path not just on the image but also on the model being explained. We introduce Adaptive Path Methods (APMs) as a generalization of path methods, and Guided IG as a specific instance of an APM. Empirically, Guided IG creates saliency maps better aligned with the model's prediction and the input image that is being explained. We show through qualitative and quantitative experiments that Guided IG outperforms other, related methods in nearly every experiment.
****
Spatiotemporal Registration for Event-Based Visual Odometry
Daqi Liu, Alvaro Parra, Tat-Jun Chin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4937-4946
A useful application of event sensing is visual odometry, especially in settings that require high-temporal resolution. The state-of-the-art method of contrast maximisation recovers the motion from a batch of events by maximising the contrast of the image of warped events. However, the cost scales with image resolution and the temporal resolution can be limited by the need for large batch sizes to yield sufficient structure in the contrast image (see supplementary material for demonstration program). In this work, we propose spatiotemporal registration as a compelling technique for event-based rotational motion estimation. We theoretically justify the approach and establish its fundamental and practical advantages over contrast maximisation. In particular, spatiotemporal registration also produces feature tracks as a by-product, which directly supports an efficient visual odometry pipeline with graph-based optimisation for motion averaging. The simplicity of our visual odometry pipeline allows it to process more than 1 M events/second. We also contribute a new event dataset for visual odometry, where motion sequences with large velocity variations were acquired using a high-precision robot arm. Our dataset will be published after the reviewing period.
****
Temporal Action Segmentation From Timestamp Supervision
Zhe Li, Yazan Abu Farha, Jurgen Gall; Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8365-8374
Temporal action segmentation approaches have been very successful recently. However, annotating videos with frame-wise labels to train such models is very expensive and time consuming. While weakly supervised methods trained using only ordered action lists require less annotation effort, the performance is still worse than fully supervised approaches. In this paper, we propose to use timestamp supervision for the temporal action segmentation task. Timestamps require a comparable annotation effort to weakly supervised approaches, and yet provide a more supervisory signal. To demonstrate the effectiveness of timestamp supervision, we propose an approach to train a segmentation model using only timestamps annotations. Our approach uses the model output and the annotated timestamps to generate frame-wise labels by detecting the action changes. We further introduce a confidence loss that forces the predicted probabilities to monotonically decrease as the distance to the timestamps increases. This ensures that all and not only the most distinctive frames of an action are learned during training. The evaluation on four datasets shows that models trained with timestamps annotations achieve comparable performance to the fully supervised approaches.
****
Data-Free Model Extraction
Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, Nicolas Papernot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4771-4780
Current model extraction attacks assume that the adversary has access to a surrogate dataset with characteristics similar to the proprietary data used to train the victim model. This requirement precludes the use of existing model extraction techniques on valuable models, such as those trained on rare or hard to acquire datasets. In contrast, we propose data-free model extraction methods that do not require a surrogate dataset. Our approach adapts techniques from the area of data-free knowledge transfer for model extraction. As part of our study, we identify that the choice of loss is critical to ensuring that the extracted model is an accurate replica of the victim model. Furthermore, we address difficulties arising from the adversary's limited access to the victim model in a black-box setting. For example, we recover the model's logits from its probability predictions to approximate gradients. We find that the proposed data-free model extraction approach achieves high-accuracy with reasonable query complexity -- 0.99x and 0.92x the victim model accuracy on SVHN and CIFAR-10 datasets given 2M and 20M queries respectively.
****
PointAugmenting: Cross-Modal Augmentation for 3D Object Detection
Chunwei Wang, Chao Ma, Ming Zhu, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11794-11803
Camera and LiDAR are two complementary sensors for 3D object detection in the autonomous driving context. Camera provides rich texture and color cues while LiDAR specializes in relative distance sensing. The challenge of 3D object detection lies in effectively fusing 2D camera images with 3D LiDAR points. In this paper, we present a novel cross-modal 3D object detection algorithm, named PointAugmenting. On one hand, PointAugmenting decorates point clouds with corresponding point-wise CNN features extracted by pretrained 2D detection models, and then performs 3D object detection over the decorated point clouds. In comparison with highly abstract semantic segmentation scores to decorate point clouds, CNN features from detection networks adapt to object appearance variations, achieving significant improvement. On the other hand, PointAugmenting benefits from a novel cross-modal data augmentation algorithm, which consistently pastes virtual objects into images and point clouds during network training. Extensive experiments on the large-scale nuScenes and Waymo datasets demonstrate the effectiveness and efficiency of our PointAugmenting. Notably, PointAugmenting outperforms the LiDAR-only baseline detector by +6.5% mAP and achieves the new state-of-the-art results on the nuScenes leaderboard to date.
****
Learning Feature Aggregation for Deep 3D Morphable Models

Zhixiang Chen, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13164-13173
3D morphable models are widely used for the shape representation of an object class in computer vision and graphics applications. In this work, we focus on deep 3D morphable models that directly apply deep learning on 3D mesh data with a hierarchical structure to capture information at multiple scales. While great efforts have been made to design the convolution operator, how to best aggregate vertex features across hierarchical levels deserves further attention. In contrast to resorting to mesh decimation, we propose an attention based module to learn mapping matrices for better feature aggregation across hierarchical levels. Specifically, the mapping matrices are generated by a compatibility function of the keys and queries. The keys and queries are trainable variables, learned by optimizing the target objective, and shared by all data samples of the same object class. Our proposed module can be used as a train-only drop-in replacement for the feature aggregation in existing architectures for both downsampling and upsampling. Our experiments show that through the end-to-end training of the mapping matrices, we achieve state-of-the-art results on a variety of 3D shape datasets in comparison to existing morphable models.
****
There Is More Than Meets the Eye: Self-Supervised Multi-Object Detection and Tracking With Sound by Distilling Multimodal Knowledge
Francisco Rivera Valverde, Juana Valeria Hurtado, Abhinav Valada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11612-11621
Attributes of sound inherent to objects can provide valuable cues to learn rich representations for object detection and tracking. Furthermore, the co-occurrence of audiovisual events in videos can be exploited to localize objects over the image field by solely monitoring the sound in the environment. Thus far, this has only been feasible in scenarios where the camera is static and for single object detection. Moreover, the robustness of these methods has been limited as they primarily rely on RGB images which are highly susceptible to illumination and weather changes. In this work, we present the novel self-supervised MM-DistillNet framework consisting of multiple teachers that leverage diverse modalities including RGB, depth, and thermal images, to simultaneously exploit complementary cues and distill knowledge into a single audio student network. We propose the new MTA loss function that facilitates the distillation of information from multimodal teachers in a self-supervised manner. Additionally, we propose a novel self-supervised pretext task for the audio student that enables us to not rely on labor-intensive manual annotations. We introduce a large-scale multimodal dataset with over 113,000 time-synchronized frames of RGB, depth, thermal, and audio modalities. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods while being able to detect multiple objects using only sound during inference and even while moving.
****
DeRF: Decomposed Radiance Fields
Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14153-14161
With the advent of Neural Radiance Fields (NeRF), neural networks can now render novel views of a 3D scene with quality that fools the human eye. Yet, generating these images is very computationally intensive, limiting their applicability in practical scenarios. In this paper, we propose a technique based on spatial decomposition capable of mitigating this issue. Our key observation is that there are diminishing returns in employing larger (deeper and/or wider) networks. Hence, we propose to spatially decompose a scene and dedicate smaller networks for each decomposed part. When working together, these networks can render the whole scene. This allows us near-constant inference time regardless of the number of decomposed parts. Moreover, we show that a Voronoi spatial decomposition is preferable for this purpose, as it is provably compatible with the Painter's Algorithm for efficient and GPU-friendly rendering. Our experiments show that for real-w

orld scenes, our method provides up to 3x more efficient inference than NeRF (wi
th the same rendering quality), or an improvement of up to 1.0 dB in PSNR (for t
he same inference cost).
****
Group-aware Label Transfer for Domain Adaptive Person Re-identification
Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, Zheng-Jun Zha; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2021, pp. 5310-5319
Unsupervised Domain Adaptive (UDA) person re-identification (ReID) aims at adapt
ing the model trained on a labeled source-domain dataset to a target-domain data
set without any further annotations. Most successful UDA-ReID approaches combine
 clustering-based pseudo-label prediction with representation learning and perfo
rm the two steps in an alternating fashion. However, offline interaction between
 these two steps may allow noisy pseudo labels to substantially hinder the capab
ility of the model. In this paper, we propose a Group-aware Label Transfer (GLT)
 algorithm, which enables the online interaction and mutual promotion of pseudo-
label prediction and representation learning. Specifically, a label transfer alg
orithm simultaneously uses pseudo labels to train the data while refining the ps
eudo labels as an online clustering algorithm. It treats the online label refine
ry problem as an optimal transport problem, which explores the minimum cost for
assigning M samples to N pseudo labels. More importantly, we introduce a group-a
ware strategy to assign implicit attribute group IDs to samples. The combination
 of the online label refining algorithm and the group-aware strategy can better
correct the noisy pseudo label in an online fashion and narrow down the search s
pace of the target identity. The effectiveness of the proposed GLT is demonstrat
ed by the experimental results (Rank-1 accuracy) for Market1501\toDukeMTMC (82.0
%) and DukeMTMC\toMarket1501 (92.2%), remarkably closing the gap between unsuper
vised and supervised performance on person re-identification.
****
MR Image Super-Resolution With Squeeze and Excitation Reasoning Attention Networ
k
Yulun Zhang, Kai Li, Kunpeng Li, Yun Fu; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13425-13434
High-quality high-resolution (HR) magnetic resonance (MR) images afford more det
ailed information for reliable diagnosis and quantitative image analyses. Deep c
onvolutional neural networks (CNNs) have shown promising ability for MR image su
per-resolution (SR) given low-resolution (LR) MR images. The LR MR images usuall
y share some visual characteristics: repeating patterns, relatively simpler stru
ctures, and less informative background. Most previous CNN-based SR methods trea
t the spatial pixels (including the background) equally. They also fail to sense
 the entire space of the input, which is critical for high-quality MR image SR.
To address those problems, we propose squeeze and excitation reasoning attention
 networks (SERAN) for accurate MR image SR. We propose to squeeze attention from
 global spatial information of the input and obtain global descriptors. Such glo
bal descriptors enhance the network's ability to focus on more informative regio
ns and structures in MR images. We further build relationship among those global
 descriptors and propose primitive relationship reasoning attention. The global
descriptors are further refined with the learned attention. To fully make use of
 the aggregated information, we adaptively recalibrate feature responses with le
arned adaptive attention vectors. These attention vectors select a subset of glo
bal descriptors to complement each spatial location for accurate details and tex
ture reconstruction. We propose squeeze and excitation attention with residual s
caling, which not only stabilizes the training but also makes it flexible to oth
er basic networks. Extensive experiments show the effectiveness of our proposed
SERAN, which clearly surpasses state-of-the-art methods on benchmarks quantitati
vely and qualitatively.
****
BABEL: Bodies, Action and Behavior With English Labels
Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros
-Ramirez, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer V

ision and Pattern Recognition (CVPR), 2021, pp. 722-731
Understanding the semantics of human movement -- the what, how and why of the mo vement -- is an important problem that requires datasets of human actions with s emantic labels. Existing datasets take one of two approaches. Large-scale video datasets contain many action labels but do not contain ground-truth 3D human mot ion. Alternatively, motion-capture (mocap) datasets have precise body motions bu t are limited to a small number of actions. To address this, we present BABEL, a large dataset with language labels describing the actions being performed in mo cap sequences. BABEL consists of language labels for over 43 hours of mocap sequ ences from AMASS, containing over 250 unique actions. Each action label in BABEL is precisely aligned with the duration of the corresponding action in the mocap sequence. BABEL also allows overlap of multiple actions, that may each span dif ferent durations. This results in a total of over 66000 action segments. The den se annotations can be leveraged for tasks like action recognition, temporal loca lization, motion synthesis, etc. To demonstrate the value of BABEL as a benchmar k, we evaluate the performance of models on 3D action recognition. We demonstrat e that BABEL poses interesting learning challenges that are applicable to real-w orld scenarios, and can serve as a useful benchmark for progress in 3D action re cognition. The dataset, baseline methods, and evaluation code are available and supported for academic research purposes at https://babel.is.tue.mpg.de/.
****
SMD-Nets: Stereo Mixture Density Networks
Fabio Tosi, Yiyi Liao, Carolin Schmitt, Andreas Geiger; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8942 -8952
Despite stereo matching accuracy has greatly improved by deep learning in the la st few years, recovering sharp boundaries and high-resolution outputs efficientl y remains challenging. In this paper, we propose Stereo Mixture Density Networks (SMD-Nets), a simple yet effective learning framework compatible with a wide cl ass of 2D and 3D architectures which ameliorates both issues. Specifically, we e xploit bimodal mixture densities as output representation and show that this all ows for sharp and precise disparity estimates near discontinuities while explici tly modeling the aleatoric uncertainty inherent in the observations. Moreover, w e formulate disparity estimation as a continuous problem in the image domain, al lowing our model to query disparities at arbitrary spatial precision. We carry o ut comprehensive experiments on a new high-resolution and highly realistic synth etic stereo dataset, consisting of stereo pairs at 8Mpx resolution, as well as o n real-world stereo datasets. Our experiments demonstrate increased depth accura cy near object boundaries and prediction of ultra high-resolution disparity maps on standard GPUs. We demonstrate the flexibility of our technique by improving the performance of a variety of stereo backbones.
****
Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification
Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Binen g Zhong, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4330-4339
Visible-infrared person re-identification (Re-ID) aims to match the pedestrian i mages of the same identity from different modalities. Existing works mainly focu s on alleviating the modality discrepancy by aligning the distributions of featu res from different modalities. However, nuanced but discriminative information, such as glasses, shoes, and the length of clothes, has not been fully explored, especially in the infrared modality. Without discovering nuances, it is challeng ing to match pedestrians across modalities using modality alignment solely, whic h inevitably reduces feature distinctiveness. In this paper, we propose a joint Modality and Pattern Alignment Network (MPANet) to discover cross-modality nuanc es in different patterns for visible-infrared person Re-ID, which introduces a m odality alleviation module and a pattern alignment module to jointly extract dis criminative features. Specifically, we first propose a modality alleviation modu le to dislodge the modality information from the extracted feature maps. Then, W e devise a pattern alignment module, which generates multiple pattern maps for t

he diverse patterns of a person, to discover nuances. Finally, we introduce a mutual mean learning fashion to alleviate the modality discrepancy and propose a center cluster loss to guide both identity learning and nuances discovering. Extensive experiments on the public SYSU-MM01 and RegDB datasets demonstrate the superiority of MPANet over state-of-the-arts.
****
Learning Progressive Point Embeddings for 3D Point Cloud Generation
Cheng Wen, Baosheng Yu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10266-10275
Generative models for 3D point clouds are extremely important for scene/object reconstruction applications in autonomous driving and robotics. Despite recent success of deep learning-based representation learning, it remains a great challenge for deep neural networks to synthesize or reconstruct high-fidelity point clouds, because of the difficulties in 1) learning effective pointwise representations; and 2) generating realistic point clouds from complex distributions. In this paper, we devise a dual-generators framework for point cloud generation, which generalizes vanilla generative adversarial learning framework in a progressive manner. Specifically, the first generator aims to learn effective point embeddings in a breadth-first manner, while the second generator is used to refine the generated point cloud based on a depth-first point embedding to generate a robust and uniform point cloud. The proposed dual-generators framework thus is able to progressively learn effective point embeddings for accurate point cloud generation. Experimental results on a variety of object categories from the most popular point cloud generation dataset, ShapeNet, demonstrate the state-of-the-art performance of the proposed method for accurate point cloud generation.
****
Learnable Graph Matching: Incorporating Graph Partitioning With Deep Feature Learning for Multiple Object Tracking
Jiawei He, Zehao Huang, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5299-5309
Data association across frames is at the core of Multiple Object Tracking (MOT) task. This problem is usually solved by a traditional graph-based optimization or directly learned via deep learning. Despite their popularity, we find some points worth studying in current paradigm: 1) Existing methods mostly ignore the context information among tracklets and intra-frame detections, which makes the tracker hard to survive in challenging cases like severe occlusion. 2) The end-to-end association methods solely rely on the data fitting power of deep neural networks, while they hardly utilize the advantage of optimization-based assignment methods. 3) The graph-based optimization methods mostly utilize a separate neural network to extract features, which brings the inconsistency between training and inference. Therefore, in this paper we propose a novel learnable graph matching method to address these issues. Briefly speaking, we model the relationships between tracklets and the intra-frame detections as a general undirected graph. Then the association problem turns into a general graph matching between tracklet graph and detection graph. Furthermore, to make the optimization end-to-end differentiable, we relax the original graph matching into continuous quadratic programming and then incorporate the training of it into a deep graph network with the help of the implicit function theorem. Lastly, our method GMTracker, achieves state-of-the-art performance on several standard MOT datasets. Our code is available at https://github.com/jiaweihe1996/GMTracker.
****
A Decomposition Model for Stereo Matching
Chengtang Yao, Yunde Jia, Huijun Di, Pengxiang Li, Yuwei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6091-6100
In this paper, we present a decomposition model for stereo matching to solve the problem of excessive growth in computational cost (time and memory cost) as the resolution increases. In order to reduce the huge cost of stereo matching at the original resolution, our model only runs dense matching at a very low resoluti

on and uses sparse matching at different higher resolutions to recover the dispa rity of lost details scale-by-scale. After the decomposition of stereo matching, our model iteratively fuses the sparse and dense disparity maps from adjacent s cales with an occlusion-aware mask. A refinement network is also applied to impr oving the fusion result. Compared with high-performance methods like PSMNet and GANet, our method achieves 10-100x speed increase while obtaining comparable dis parity estimation results.
****
RangeIoUDet: Range Image Based Real-Time 3D Object Detector Optimized by Interse ction Over Union
Zhidong Liang, Zehan Zhang, Ming Zhang, Xian Zhao, Shiliang Pu; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7140-7149
Real-time and high-performance 3D object detection is an attractive research dir ection in autonomous driving. Recent studies prefer point based or voxel based c onvolution for achieving high performance. However, these methods suffer from th e unsatisfied efficiency or complex customized convolution, making them unsuitab le for applications with real-time requirements. In this paper, we present an ef ficient and effective 3D object detection framework, named RangeIoUDet that uses the range image as input. Benefiting from the dense representation of the range image, RangeIoUDet is entirely constructed based on 2D convolution, making it p ossible to have a fast inference speed. This model learns pointwise features fro m the range image, which is then passed to a region proposal network for predict ing 3D bounding boxes. We optimize the pointwise feature and the 3D box via the point-based IoU and box-based IoU supervision, respectively. The point-based IoU supervision is proposed to make the network better learn the implicit 3D inform ation encoded in the range image. The 3D Hybrid GIoU loss is introduced to gener ate high-quality boxes while providing an accurate quality evaluation. Through t he point-based IoU and the box-based IoU, RangeIoUDet outperforms all single-sta ge models on the KITTI dataset, while running at 45 FPS for inference. Experimen ts on the self-built dataset further prove its effectiveness on different LIDAR sensors and object categories.
****
Domain-Robust VQA With Diverse Datasets and Methods but No Target Labels
Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, Rebecca Hwa; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2021, pp. 7046-7056
The observation that computer vision methods overfit to dataset specifics has in spired diverse attempts to make object recognition models robust to domain shift s. However, similar work on domain-robust visual question answering methods is v ery limited. Domain adaptation for VQA differs from adaptation for object recogn ition due to additional complexity: VQA models handle multimodal inputs, methods contain multiple steps with diverse modules resulting in complex optimization, and answer spaces in different datasets are vastly different. To tackle these ch allenges, we first quantify domain shifts between popular VQA datasets, in both visual and textual space. To disentangle shifts between datasets arising from di fferent modalities, we also construct synthetic shifts in the image and question domains separately. Second, we test the robustness of different families of VQA methods (classic two-stream, transformer, and neuro-symbolic methods) to these shifts. Third, we test the applicability of existing domain adaptation methods a nd devise a new one to bridge VQA domain gaps, adjusted to specific VQA models. To emulate the setting of real-world generalization, we focus on unsupervised do main adaptation and the open-ended classification task formulation.
****
(AF)2-S3Net: Attentive Feature Fusion With Adaptive Feature Selection for Sparse Semantic Segmentation Network
Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, Bingbing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp . 12547-12556
Autonomous robotic systems and self driving cars rely on accurate perception of

their surroundings as the safety of the passengers and pedestrians is the top priority. Semantic segmentation is one the essential components of environmental perception that provides semantic information of the scene. Recently, several methods have been introduced for 3D LiDAR semantic segmentation. While, they can lead to improved performance, they are either afflicted by high computational complexity, therefore are inefficient, or lack fine details of smaller instances. To alleviate this problem, we propose AF2-S3Net, an end-to-end encoder-decoder CNN network for 3D LiDAR semantic segmentation. We present a novel multi-branch attentive feature fusion module in the encoder and a unique adaptive feature selection module with feature map re-weighting in the decoder. Our AF2-S3Net fuses the voxel based learning and point-based learning into a single framework to effectively process the large 3D scene. Our experimental results show that the proposed method outperforms the state-of-the-art approaches on the large-scale nuScenes-lidarseg and SemanticKITTI benchmark, ranking 1st on both competitive public leaderboard competitions upon publication.
****
Towards Real-World Blind Face Restoration With Generative Facial Prior
Xintao Wang, Yu Li, Honglun Zhang, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9168-9178
Blind face restoration usually relies on facial priors, such as facial geometry prior or reference prior, to restore realistic and faithful details. However, very low-quality inputs cannot offer accurate geometric prior while high-quality references are inaccessible, limiting the applicability in real-world scenarios. In this work, we propose GFP-GAN that leverages rich and diverse priors encapsulated in a pretrained face GAN for blind face restoration. This Generative Facial Prior (GFP) is incorporated into the face restoration process via spatial feature transform layers, which allow our method to achieve a good balance of realness and fidelity. Thanks to the powerful generative facial prior and delicate designs, our GFP-GAN could jointly restore facial details and enhance colors with just a single forward pass, while GAN inversion methods require image-specific optimization at inference. Extensive experiments show that our method achieves superior performance to prior art on both synthetic and real-world datasets.
****
Track To Detect and Segment: An Online Multi-Object Tracker
Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12352-12361
Most online multi-object trackers perform object detection stand-alone in a neural net without any input from tracking. In this paper, we present a new online joint detection and tracking model, TraDeS (TRAck to DEtect and Segment), exploiting tracking clues to assist detection end-to-end. TraDeS infers object tracking offset by a cost volume, which is used to propagate previous object features for improving current object detection and segmentation. Effectiveness and superiority of TraDeS are shown on 4 datasets, including MOT (2D tracking), nuScenes (3D tracking), MOTS and Youtube-VIS (instance segmentation tracking). Project page: https://jialianwu.com/projects/TraDeS.html.
****
Look Before You Speak: Visually Contextualized Utterances
Paul Hongsuck Seo, Arsha Nagrani, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16877-16887
While most conversational AI systems focus on textual dialogue only, conditioning utterances on visual context (when it's available) can lead to more realistic conversations. Unfortunately, a major challenge for incorporating visual context into conversational dialogue is the lack of large-scale labeled datasets. We provide a solution in the form of a new visually conditioned Future Utterance Prediction task. Our task involves predicting the next utterance in a video, using both visual frames and transcribed speech as context. By exploiting the large number of instructional videos online, we train a model to solve this task at scale, without the need for manual annotations. Leveraging recent advances in multimo

dal learning, our model consists of a novel co-attentional multimodal video tran
sformer, and when trained on both textual and visual context, outperforms baseli
nes that use textual inputs alone. Further, we demonstrate that our model traine
d for this task on unlabelled videos achieves state-of-the-art performance on a
number of downstream VideoQA benchmarks such as MSRVTT-QA, MSVD-QA, ActivityNet-
QA and How2QA.
****
DivCo: Diverse Conditional Image Synthesis via Contrastive Generative Adversaria
l Network
Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, Hongsheng Li; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
 pp. 16377-16386
Conditional generative adversarial networks (cGANs) target at synthesizing diver
se images given the input conditions and latent codes, but unfortunately, they u
sually suffer from the issue of mode collapse. Towards solving this issue, previ
ous works mainly focused on encouraging the correlation between the latent codes
 and the generated images, while ignoring the relations between images generated
 from various latent codes. The recent MSGAN tried to encourage the diversity of
 the generated image but still only considers "negative" relations between the i
mage pairs. In this paper, we propose a novel DivCo framework to properly constr
ain both "positive" and "negative" relations between the generated images specif
ied in the latent space. To the best of our knowledge, this is the first attempt
 to use contrastive learning for diverse conditional image synthesis. A latent-a
ugmented contrastive loss is introduced, which encourage images generated from a
djacent latent codes to be similar and those generated from distinct latent code
s to show low affinities. The proposed latent-augmented contrastive loss are wel
l compatible with various cGAN architectures. Extensive experiments demonstrate
the proposed DivCo could produce more diverse images than state-of-the-art metho
ds without sacrificing visual quality in multiple settings.
****
Effective Sparsification of Neural Networks With Global Sparsity Constraint
Xiao Zhou, Weizhong Zhang, Hang Xu, Tong Zhang; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3599-3608
Weight pruning is an effective technique to reduce the model size and inference
time for deep neural networks in real world deployments. However, since magnitud
es and relative importance of weights are very different for different layers of
 a neural network, existing methods rely on either manual tuning or handcrafted
heuristic rules to find appropriate pruning rates individually for each layer. T
his approach general leads to suboptimal performance. In this paper, by directly
 working on the probability space, we propose an effective network sparsificatio
n method called probabilistic masking (ProbMask), which solves a natural sparsif
ication formulation under global sparsity constraint. The key idea is to use pro
bability as a global criterion for all layers to measure the weight importance.
An appealing feature of ProbMask is that the amounts of weight redundancy can be
 learned automatically via our constraint and thus we avoid the problem of tunin
g pruning rates individually for different layers in a network. Extensive experi
mental results on CIFAR-10/100 and ImageNet demonstrate that our method is highl
y effective, and can outperform previous state-of-the-art methods by a significa
nt margin, especially in the high pruning rate situation. Notably, the gap of To
p-1 accuracy between our ProbMask and existing methods can be up to 10%. As a by
-product, we show ProbMask is also highly effective in identifying supermasks, w
hich are subnetworks with high performance in a randomly weighted dense neural n
etwork.
****
Deep Gaussian Scale Mixture Prior for Spectral Compressive Imaging
Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, Guangming Shi; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p
p. 16216-16225
In coded aperture snapshot spectral imaging (CASSI) system, the real-world hyper
spectral image (HSI) can be reconstructed from the captured compressive image in

a snapshot. Model-based HSI reconstruction methods employed hand-crafted priors to solve the reconstruction problem, but most of which achieved limited success due to the poor representation capability of these hand-crafted priors. Deep learning based methods learning the mappings between the compressive images and the HSIs directly achieved much better results. Yet, it is nontrivial to design a powerful deep network heuristically for achieving satisfied results. In this paper, we propose a novel HSI reconstruction method based on the Maximum a Posterior (MAP) estimation framework using learned Gaussian Scale Mixture (GSM) prior. Different from existing GSM models using hand-crafted scale priors (e.g., the Jeffrey's prior), we propose to learn the scale prior through a deep convolutional neural network (DCNN). Furthermore, we also propose to estimate the local means of the GSM models by the DCNN. All the parameters of the MAP estimation algorithm and the DCNN parameters are jointly optimized through end-to-end training. Extensive experimental results on both synthetic and real datasets demonstrate that the proposed method outperforms existing state-of-the-art methods. The code is available at https://see.xidian.edu.cn/faculty/wsdong/Projects/DGSM-SCI.htm.
****
Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation
Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, Ke Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3937-3946
Unsupervised Domain Adaptation (UDA) aims to generalize the knowledge learned from a well-labeled source domain to an unlabled target domain. Recently, adversarial domain adaptation with two distinct classifiers (bi-classifier) has been introduced into UDA which is effective to align distributions between different domains. Previous bi-classifier adversarial learning methods only focus on the similarity between the outputs of two distinct classifiers. However, the similarity of the outputs cannot guarantee the accuracy of target samples, i.e., traget samples may match to wrong categories even if the discrepancy between two classifiers is small. To challenge this issue, in this paper, we propose a cross-domain gradient discrepancy minimization (CGDM) method which explicitly minimizes the discrepancy of gradients generated by source samples and target samples. Specifically, the gradient gives a cue for the semantic information of target samples so it can be used as a good supervision to improve the accuracy of target samples. In order to compute the gradient signal of target smaples, we further obtain target pseudo labels through a clustering-based self-supervised learning. Extensive experiments on three widely used UDA datasets show that our method surpasses many previous state-of-the-arts.
****
DISCO: Dynamic and Invariant Sensitive Channel Obfuscation for Deep Neural Networks
Abhishek Singh, Ayush Chopra, Ethan Garza, Emily Zhang, Praneeth Vepakomma, Vivek Sharma, Ramesh Raskar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12125-12135
Recent deep learning models have shown remarkable performance in image classification. While these deep learning systems are getting closer to practical deployment, the common assumption made about data is that it does not carry any sensitive information. This assumption may not hold for many practical cases, especially in the domain where an individual's personal information is involved, like healthcare and facial recognition systems. We posit that selectively removing features in this latent space can protect the sensitive information and provide better privacy-utility trade-off. Consequently, we propose DISCO which learns a dynamic and data driven pruning filter to selectively obfuscate sensitive information in the feature space. We propose diverse attack schemes for sensitive inputs and attributes and demonstrate the effectiveness of DISCO against state-of-the-art methods through quantitative and qualitative evaluation. Finally, we also release an evaluation benchmark dataset of 1 million sensitive representations to encourage rigorous exploration of novel attack and defense schemes at https://github.com/splitlearning/InferenceBenchmark
****

Training Generative Adversarial Networks in One Stage
Chengchao Shen, Youtan Yin, Xinchao Wang, Xubin Li, Jie Song, Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3350-3360
Generative Adversarial Networks (GANs) have demonstrated unprecedented success in various image generation tasks. The encouraging results, however, come at the price of a cumbersome training process, during which the generator and discriminator are alternately updated in two stages. In this paper, we investigate a general training scheme that enables training GANs efficiently in only one stage. Based on the adversarial losses of the generator and discriminator, we categorize GANs into two classes, Symmetric GANs and Asymmetric GANs, and introduce a novel gradient decomposition method to unify the two, allowing us to train both classes in one stage and hence alleviate the training effort. We also computationally analyze the efficiency of the proposed method, and empirically demonstrate that, the proposed method yields a solid: 1.5x acceleration across various datasets and network architectures. Furthermore, we show that the proposed method is readily applicable to other adversarial-training scenarios, such as data-free knowledge distillation. The code is available at https://github.com/zju-vipa/OSGAN.
****
Learning To Aggregate and Personalize 3D Face From In-the-Wild Photo Collection
Zhenyu Zhang, Yanhao Ge, Renwang Chen, Ying Tai, Yan Yan, Jian Yang, Chengjie Wang, Jilin Li, Feiyue Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14214-14224
Non-prior face modeling aims to reconstruct 3D face only from images without shape assumptions. While plausible facial details are predicted, the models tend to over-depend on local color appearance and suffer from ambiguous noise. To address such problem, this paper presents a novel Learning to Aggregate and Personalize (LAP) framework for unsupervised robust 3D face modeling. Instead of using controlled environment, the proposed method implicitly disentangles ID-consistent and scene-specific face from unconstrained photo set. Specifically, to learn ID-consistent face, LAP adaptively aggregates intrinsic face factors of an identity based on a novel curriculum learning approach with relaxed consistency loss. To adapt the face for a personalized scene, we propose a novel attribute-refining network to modify ID-consistent face with target attribute and details. Based on the proposed method, we make unsupervised 3D face modeling benefit from meaningful image facial structure and possibly higher resolutions. Extensive experiments on benchmarks show LAP recovers superior or competitive face shape and texture, compared with state-of-the-art (SOTA) methods with or without prior and supervision.
****
Leveraging Line-Point Consistence To Preserve Structures for Wide Parallax Image Stitching
Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xinchen Ye, Longin Jan Latecki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12186-12195
Generating high-quality stitched images with natural structures is a challenging task in computer vision. In this paper, we succeed in preserving both local and global geometric structures for wide parallax images, while reducing artifacts and distortions. A projective invariant, Characteristic Number, is used to match co-planar local sub-regions for input images. The homography between these well-matched sub-regions produces consistent line and point pairs, suppressing artifacts in overlapping areas. We explore and introduce global collinear structures into an objective function to specify and balance the desired characters for image warping, which can preserve both local and global structures while alleviating distortions. We also develop comprehensive measures for stitching quality to quantify the collinearity of points and the discrepancy of matched line pairs by considering the sensitivity to linear structures for human vision. Extensive experiments demonstrate the superior performance of the proposed method over the state-of-the-art by presenting sharp textures and preserving prominent natural structures in stitched images. Especially, our method not only exhibits lower error

s but also the least divergence across all test images. Code is available at htt ps://github.com/dut-media-lab/Image-Stitching
****
3DIoUMatch: Leveraging IoU Prediction for Semi-Supervised 3D Object Detection
He Wang, Yezhen Cong, Or Litany, Yue Gao, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14615-14624
3D object detection is an important yet demanding task that heavily relies on difficult to obtain 3D annotations. To reduce the required amount of supervision, we propose 3DIoUMatch, a novel semi-supervised method for 3D object detection applicable to both indoor and outdoor scenes. We leverage a teacher-student mutual learning framework to propagate information from the labeled to the unlabeled train set in the form of pseudo-labels. However, due to the high task complexity, we observe that the pseudo-labels suffer from significant noise and are thus not directly usable. To that end, we introduce a confidence-based filtering mechanism, inspired by FixMatch. We set confidence thresholds based upon the predicted objectness and class probability to filter low-quality pseudo-labels. While effective, we observe that these two measures do not sufficiently capture localization quality. We therefore propose to use the estimated 3D IoU as a localization metric and set category-aware self-adjusted thresholds to filter poorly localized proposals. We adopt VoteNet as our backbone detector on indoor datasets while we use PV-RCNN on the autonomous driving dataset, KITTI. Our method consistently improves state-of-the-art methods on both ScanNet and SUN-RGBD benchmarks by significant margins under all label ratios (including fully labeled setting). For example, when training using only 10% labeled data on ScanNet, 3DIoUMatch achieves 7.7 absolute improvement on mAP@0.25 and 8.5 absolute improvement on mAP@0.5 upon the prior art. On KITTI, we are the first to demonstrate semi-supervised 3D object detection and our method surpasses a fully supervised baseline from 1.8% to 7.6% under different label ratio and categories.
****
Self-Supervised Video GANs: Learning for Appearance Consistency and Motion Coherency
Sangeek Hyun, Jihwan Kim, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10826-10835
A video can be represented by the composition of appearance and motion. Appearance (or content) expresses the information invariant throughout time, and motion describes the time-variant movement. Here, we propose self-supervised approaches for video Generative Adversarial Networks (GANs) to achieve the appearance consistency and motion coherency in videos. Specifically, the dual discriminators for image and video individually learn to solve their own pretext tasks; appearance contrastive learning and temporal structure puzzle. The proposed tasks enable the discriminators to learn representations of appearance and temporal context, and force the generator to synthesize videos with consistent appearance and natural flow of motions. Extensive experiments in facial expression and human action public benchmarks show that our method outperforms the state-of-the-art video GANs. Moreover, consistent improvements regardless of the architecture of video GANs confirm that our framework is generic.
****
Neural Lumigraph Rendering
Petr Kellnhofer, Lars C. Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4287-4297
Novel view synthesis is a challenging and ill-posed inverse rendering problem. Neural rendering techniques have recently achieved photorealistic image quality for this task. State-of-the-art (SOTA) neural volume rendering approaches, however, are slow to train and require minutes of inference (i.e., rendering) time for high image resolutions. We adopt high-capacity neural scene representations with periodic activations for jointly optimizing an implicit surface and a radiance field of a scene supervised exclusively with posed 2D images. Our neural rendering pipeline accelerates SOTA neural volume rendering by about two orders of mag

nitude and our implicit surface representation is unique in allowing us to expor
t a mesh with view-dependent texture information. Thus, like other implicit surf
ace representations, ours is compatible with traditional graphics pipelines, ena
bling real-time rendering rates, while achieving unprecedented image quality com
pared to other surface methods. We assess the quality of our approach using exis
ting datasets as well as high-quality 3D face data captured with a custom multi-
camera rig.
****
Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar a
nd Radar Signals
Kun Qian, Shilin Zhu, Xinyu Zhang, Li Erran Li; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 444-453
Vehicle detection with visual sensors like lidar and camera is one of the critic
al functions enabling autonomous driving. While they generate fine-grained point
 clouds or high-resolution images with rich information in good weather conditio
ns, they fail in adverse weather (e.g., fog) where opaque particles distort ligh
ts and significantly reduce visibility. Thus, existing methods relying on lidar
or camera experience significant performance degradation in rare but critical ad
verse weather conditions. To remedy this, we resort to exploiting complementary
radar, which is less impacted by adverse weather and becomes prevalent on vehicl
es. In this paper, we present Multimodal Vehicle Detection Network (MVDNet), a t
wo-stage deep fusion detector, which first generates proposals from two sensors
and then fuses region-wise features between multimodal sensor streams to improve
 final detection results. To evaluate MVDNet, we create a procedurally generated
 training dataset based on the collected raw lidar and radar signals from the op
en-source Oxford Radar Robotcar. We show that the proposed MVDNet surpasses othe
r state-of-the-art methods, notably in terms of Average Precision (AP), especial
ly in adverse weather conditions. The code and data are available at https://git
hub.com/qiank10/MVDNet.
****
Stochastic Whitening Batch Normalization
Shengdong Zhang, Ehsan Nezhadarya, Homa Fashandi, Jiayi Liu, Darin Graham, Mohak
 Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2021, pp. 10978-10987
Batch Normalization (BN) is a popular technique for training Deep Neural Network
s (DNNs). BN uses scaling and shifting to normalize activations of mini-batches
to accelerate convergence and improve generalization. The recently proposed Iter
ative Normalization (IterNorm) method improves these properties by whitening the
 activations iteratively using Newton's method. However, since Newton's method i
nitializes the whitening matrix independently at each training step, no informat
ion is shared between consecutive steps. In this work, instead of exact computat
ion of whitening matrix at each time step, we estimate it gradually during train
ing in an online fashion, using our proposed Stochastic Whitening Batch Normaliz
ation (SWBN) algorithm. We show that while SWBN improves the convergence rate an
d generalization of DNNs, its computational overhead is less than that of IterNo
rm. Due to the high efficiency of the proposed method, it can be easily employed
 in most DNN architectures with a large number of layers. We provide comprehensi
ve experiments and comparisons between BN, IterNorm, and SWBN layers to demonstr
ate the effectiveness of the proposed technique in conventional (many-shot) imag
e classification and few-shot classification tasks.
****
Self-Guided and Cross-Guided Learning for Few-Shot Segmentation
Bingfeng Zhang, Jimin Xiao, Terry Qin; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8312-8321
Few-shot segmentation has been attracting a lot of attention due to its effectiv
eness to segment unseen object classes with a few annotated samples. Most existi
ng approaches use masked Global Average Pooling (GAP) to encode an annotated sup
port image to a feature vector to facilitate query image segmentation. However,
this pipeline unavoidably loses some discriminative information due to the avera
ge operation. In this paper, we propose a simple but effective self-guided learn

ing approach, where the lost critical information is mined. Specifically, throug h making an initial prediction for the annotated support image, the covered and uncovered foreground regions are encoded to the primary and auxiliary support ve ctors using masked GAP, respectively. By aggregating both the primary and auxili ary support vectors, better segmentation performance is obtained on query images . Enlightened by our self-guided module for 1-shot segmentation, we propose a cr oss-guided module for multiple shot segmentation, where the final mask is fused using predictions from multiple annotated samples with high-quality support vect ors contributing more and vice versa. This module improves the final prediction in the inference stage without re-training. Extensive experiments show that our approach achieves new state-of-the-art performances on both PASCAL-5i and COCO-2 0i datasets. Source code will be released once the paper is accepted.
****
M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pr e-Training
Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongd ong Zhang, Nan Duan; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2021, pp. 3977-3986
We present M3P, a Multitask Multilingual Multimodal Pre-trained model that combi nes multilingual pre-training and multimodal pre-training into a unified framewo rk via multitask pre-training. Our goal is to learn universal representations th at can map objects occurred in different modalities or texts expressed in differ ent languages into a common semantic space. In addition, to explicitly encourage fine-grained alignment between images and non-English languages, we also propos e Multimodal Code-switched Training (MCT) to combine monolingual pre-training an d multimodal pre-training via a code-switch strategy. Experiments are performed on the multilingual image retrieval task across two benchmark datasets, includin g MSCOCO and Multi30K. M3P can achieve comparable results for English and new st ate-of-the-art results for non-English languages.
****
Hyperdimensional Computing as a Framework for Systematic Aggregation of Image De scriptors
Peer Neubert, Stefan Schubert; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2021, pp. 16938-16947
Image and video descriptors are an omnipresent tool in computer vision and its a pplication fields like mobile robotics. Many hand-crafted and in particular lear ned image descriptors are numerical vectors with a potentially (very) large numb er of dimensions. Practical considerations like memory consumption or time for c omparisons call for the creation of compact representations. In this paper, we u se hyperdimensional computing (HDC) as an approach to systematically combine inf ormation from a set of vectors in a single vector of the same dimensionality. HD C is a known technique to perform symbolic processing with distributed represent ations in numerical vectors with thousands of dimensions. We present a HDC imple mentation that is suitable for processing the output of existing and future (dee p learning based) image descriptors. We discuss how this can be used as a framew ork to process descriptors together with additional knowledge by simple and fast vector operations. A concrete outcome is a novel HDC-based approach to aggregat e a set of local image descriptors together with their image positions in a sing le holistic descriptor. The comparison to available holistic descriptors and agg regation methods on a series of standard mobile robotics place recognition exper iments shows a 20% improvement in average performance and >2x better worst-case performance compared to runner-up.
****
Layerwise Optimization by Gradient Decomposition for Continual Learning
Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9634-9643
Deep neural networks achieve state-of-the-art and sometimes super-human performa nce across a variety of domains. However, when learning tasks sequentially, the networks easily forget the knowledge of previous tasks, known as "catastrophic f

orgetting". To achieve the consistencies between the old tasks and the new task, one effective solution is to modify the gradient for update. Previous methods enforce independent gradient constraints for different tasks, while we consider these gradients contain complex information, and propose to leverage inter-task information by gradient decomposition. In particular, the gradient of an old task is decomposed into a part shared by all old tasks and a part specific to that task. The gradient for update should be close to the gradient of the new task, consistent with the gradients shared by all old tasks, and orthogonal to the space spanned by the gradients specific to the old tasks. In this way, our approach will encourage common knowledge consolidation but will not impair the task-specific knowledge. Furthermore, the optimization is performed for the gradients of each layer separately rather than the concatenation of all gradients as in previous works. This effectively avoids the influence of the magnitude variation of the gradients in different layers. Extensive experiments validate the effectiveness of both gradient-decomposed optimization and layer-wise updates. Our proposed method achieves state-of-the-art results on various benchmarks of continual learning.
****
Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging
S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, Yagiz Aksoy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9685-9694
Neural networks have shown great abilities in estimating depth from a single image. However, the inferred depth maps are well below one-megapixel resolution and often lack fine-grained details, which limits their practicality. Our method builds on our analysis on how the input resolution and the scene structure affects depth estimation performance. We demonstrate that there is a trade-off between a consistent scene structure and the high-frequency details, and merge low- and high-resolution estimations to take advantage of this duality using a simple depth merging network. We present a double estimation method that improves the whole-image depth estimation and a patch selection method that adds local details to the final result. We demonstrate that by merging estimations at different resolutions with changing context, we can generate multi-megapixel depth maps with a high level of detail using a pre-trained model.
****
Blind Deblurring for Saturated Images
Liang Chen, Jiawei Zhang, Songnan Lin, Faming Fang, Jimmy S. Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6308-6316
Blind deblurring has received considerable attention in recent years. However, state-of-the-art methods often fail to process saturated blurry images. The main reason is that saturated pixels are not conforming to the commonly used linear blur model. Pioneer arts suggest excluding saturated pixels during the deblurring process, which sacrifices the informative edges from saturated regions and results in insufficient information for kernel estimation when large saturated regions exist. To address this problem, we introduce a new blur model to fit both saturated and unsaturated pixels, and all informative pixels can be considered during deblurring process. Based on our model, we develop an effective maximum a posterior (MAP)-based optimization framework. Quantitative and qualitative evaluations on benchmark datasets and challenging real-world examples show that the proposed method performs favorably against existing methods.
****
Turning Frequency to Resolution: Video Super-Resolution via Event Cameras
Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7772-7781
State-of-the-art video super-resolution (VSR) methods focus on exploiting inter- and intra-frame correlations to estimate high-resolution (HR) video frames from low-resolution (LR) ones. In this paper, we study VSR from an exotic perspectiv

e, by explicitly looking into the role of temporal frequency of video frames. Th rough experiments, we observe that a higher frequency, and hence a smaller pixel displacement between consecutive frames, tends to deliver favorable super-resol ved results. This discovery motivates us to introduce Event Cameras, a novel sen sing device that responds instantly to pixel intensity changes and produces up t o millions of asynchronous events per second, to facilitate VSR. To this end, we propose an Event-based VSR framework (E-VSR), of which the key component is an asynchronous interpolation (EAI) module that reconstructs a high-frequency (HF) video stream with uniform and tiny pixel displacements between neighboring frame s from an event stream. The derived HF video stream is then encoded into a VSR m odule to recover the desired HR videos. Furthermore, an LR bi-directional interp olation loss and an HR self-supervision loss are also introduced to respectively regulate the EAI and VSR modules. Experiments on both real-world and synthetic datasets demonstrate that the proposed approach yields results superior to the s tate of the art.
****
Time Adaptive Recurrent Neural Network
Anil Kag, Venkatesh Saligrama; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2021, pp. 15149-15158
We propose a learning method that, dynamically modifies the time-constants of th e continuous-time counterpart of a vanilla RNN. The time-constants are modified based on the current observation and hidden state. Our proposal overcomes the is sues of RNN trainability, by mitigating exploding and vanishing gradient phenome na based on placing novel constraints on the parameter space, and by suppressing noise in inputs based on pondering over informative inputs to strengthen their contribution in the hidden state. As a result, our method is computationally eff icient overcoming overheads of many existing methods that also attempt to improv e RNN training. Our RNNs, despite being simpler and having light memory footprin t, shows competitive performance against standard LSTMs and baseline RNN models on many benchmark datasets including those that require long-term memory.
****
DeFMO: Deblurring and Shape Recovery of Fast Moving Objects
Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, Jiri Matas, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2021, pp. 3456-3465
Objects moving at high speed appear significantly blurred when captured with cam eras. The blurry appearance is especially ambiguous when the object has complex shape or texture. In such cases, classical methods, or even humans, are unable t o recover the object's appearance and motion. We propose a method that, given a single image with its estimated background, outputs the object's appearance and position in a series of sub-frames as if captured by a high-speed camera (i.e. t emporal super-resolution). The proposed generative model embeds an image of the blurred object into a latent space representation, disentangles the background, and renders the sharp appearance. Inspired by the image formation model, we desi gn novel self-supervised loss function terms that boost performance and show goo d generalization capabilities. The proposed DeFMO method is trained on a complex synthetic dataset, yet it performs well on real-world data from several dataset s. DeFMO outperforms the state of the art and generates high-quality temporal su per-resolution frames.
****
PISE: Person Image Synthesis and Editing With Decoupled GAN
Jinsong Zhang, Kun Li, Yu-Kun Lai, Jingyu Yang; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7982-7990
Person image synthesis, e.g., pose transfer, is a challenging problem due to lar ge variation and occlusion. Existing methods have difficulties predicting reason able invisible regions and fail to decouple the shape and style of clothing, whi ch limits their applications on person image editing. In this paper, we propose PISE, a novel two-stage generative model for person image synthesis and editing, which can generate realistic person images with desired poses, textures, and se mantic layouts. To better predict the invisible region, we first synthesize a hu

man parsing map aligned with the target pose to represent the shape of clothing by a parsing generator, and then generate the final image by an image generator. To decouple the shape and style of clothing, we propose joint global and local per-region encoding and normalization to predict the reasonable style of clothing for invisible regions. We also propose spatial-aware normalization to retain the spatial context relationship in the source image. The results of qualitative and quantitative experiments demonstrate the superiority of our model. Besides, the results of texture transfer and parsing editing show that our model can be applied to person image editing.
****
4D Hyperspectral Photoacoustic Data Restoration With Reliability Analysis
Weihang Liao, Art Subpa-asa, Yinqiang Zheng, Imari Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4598-4607
Hyperspectral photoacoustic (HSPA) spectroscopy is an emerging bi-modal imaging technology that is able to show the wavelength-dependent absorption distribution of the interior of a 3D volume. However, HSPA devices have to scan an object exhaustively in the spatial and spectral domains; and the acquired data tend to suffer from complex noise. This time-consuming scanning process and noise severely affects the usability of HSPA. It is therefore critical to examine the feasibility of 4D HSPA data restoration from an incomplete and noisy observation. In this work, we present a data reliability analysis for the depth and spectral domain. On the basis of this analysis, we explore the inherent data correlations and develop a restoration algorithm to recover 4D HSPA cubes. Experiments on real data verify that the proposed method achieves satisfactory restoration results.
****
OBoW: Online Bag-of-Visual-Words Generation for Self-Supervised Learning
Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, Patrick Perez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6830-6840
Learning image representations without human supervision is an important and active research field. Several recent approaches have successfully leveraged the idea of making such a representation invariant under different types of perturbations, especially via contrastive-based instance discrimination training. Although effective visual representations should indeed exhibit such invariances, there are other important characteristics, such as encoding contextual reasoning skills, for which alternative reconstruction-based approaches might be better suited. With this in mind, we propose a teacher-student scheme to learn representations by training a convolutional net to reconstruct a bag-of-visual-words (BoW) representation of an image, given as input a perturbed version of that same image. Our strategy performs an online training of both the teacher network (whose role is to generate the BoW targets) and the student network (whose role is to learn representations), along with an online update of the visual-words vocabulary (used for the BoW targets). This idea effectively enables fully online BoW-guided unsupervised learning. Extensive experiments demonstrate the interest of our BoW-based strategy, which surpasses previous state-of-the-art methods (including contrastive-based ones) in several applications. For instance, in downstream tasks such Pascal object detection, Pascal classification and Places205 classification, our method improves over all prior unsupervised approaches, thus establishing new state-of-the-art results that are also significantly better even than those of supervised pre-training. We provide the implementation code at https://github.com/valeoai/obow.
****
Learning-Based Image Registration With Meta-Regularization
Ebrahim Al Safadi, Xubo Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10928-10937
We introduce a meta-regularization framework for learning-based image registration. Current learning-based image registration methods use high-resolution architectures such as U-Nets to produce spatial transformations, and impose simple and explicit regularization on the output of the network to ensure that the estimat

ed displacements are smooth. While this approach works well on small deformations, it has been known to struggle when the deformations are large. Our method uses a more advanced form of meta-regularization to increase the generalization ability of learned registration models. We motivate our approach based on Reproducing Kernel Hilbert Space (RKHS) theory, and approximate that framework via a meta-regularization convolutional layer with radially symmetric, positive semi-definite filters that inherent its regularization properties. We then provide a method to learn such regularization filters while also learning to register. Our experiments on synthetic and real datasets as well as ablation analysis show that our method can improve anatomical correspondence compared to competing methods, and reduce the percentage of folding and tear in the large deformation setting, reflecting better regularization and model generalization.
****

A Hyperbolic-to-Hyperbolic Graph Convolutional Network
Jindou Dai, Yuwei Wu, Zhi Gao, Yunde Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 154-163
Hyperbolic graph convolutional networks (GCNs) demonstrate powerful representation ability to model graphs with hierarchical structure. Existing hyperbolic GCNs resort to tangent spaces to realize graph convolution on hyperbolic manifolds, which is inferior because tangent space is only a local approximation of a manifold. In this paper, we propose a hyperbolic-to-hyperbolic graph convolutional network (H2H-GCN) that directly works on hyperbolic manifolds. Specifically, we developed a manifold-preserving graph convolution that consists of a hyperbolic feature transformation and a hyperbolic neighborhood aggregation. The hyperbolic feature transformation works as linear transformation on hyperbolic manifolds. It ensures the transformed node representations still lie on the hyperbolic manifold by imposing the orthogonal constraint on the transformation sub-matrix. The hyperbolic neighborhood aggregation updates each node representation via the Einstein midpoint. The H2H-GCN avoids the distortion caused by tangent space approximations and keeps the global hyperbolic structure. Extensive experiments show that the H2H-GCN achieves substantial improvements on the link prediction, node classification, and graph classification tasks.
****

Deep Homography for Efficient Stereo Image Compression
Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1492-1501
In this paper, we propose HESIC, an end-to-end trainable deep network for stereo image compression (SIC). To fully explore the mutual information across two stereo images, we use a deep regression model to estimate the homography matrix, i.e., H matrix. Then, the left image is spatially transformed by the H matrix, and only the residual information between the left and right images is encoded to save bit-rates. A two-branch auto-encoder architecture is adopted in HESIC, corresponding to the left and right images, respectively. For entropy coding, we propose two conditional stereo entropy models, i.e., Gaussian mixture model (GMM) based and context based entropy models, to fully explore the correlation between the two images to reduce the coding bit-rates. In decoding, a cross quality enhancement module is proposed to enhance the image quality based on inverse H matrix. Experimental results show that our HESIC outperforms state-of-the-art SIC methods on InStereo2K and KITTI datasets both quantitatively and qualitatively.
****

Point2Skeleton: Learning Skeletal Representations from Point Clouds
Cheng Lin, Changjian Li, Yuan Liu, Nenglun Chen, Yi-King Choi, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4277-4286
We introduce Point2Skeleton, an unsupervised method to learn skeletal representations from point clouds. Existing skeletonization methods are limited to tubular shapes and the stringent requirement of watertight input, while our method aims to produce more generalized skeletal representations for complex structures and handle point clouds. Our key idea is to use the insights of the medial axis tra

nsform (MAT) to capture the intrinsic geometric and topological natures of the o riginal input points. We first predict a set of skeletal points by learning a ge ometric transformation, and then analyze the connectivity of the skeletal points to form skeletal mesh structures. Extensive evaluations and comparisons show ou r method has superior performance and robustness. The learned skeletal represent ation will benefit several unsupervised tasks for point clouds, such as surface reconstruction and segmentation.
****
Neighborhood Contrastive Learning for Novel Class Discovery
Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, Nicu Sebe; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10867-10875
In this paper, we address Novel Class Discovery (NCD), the task of unveiling new classes in a set of unlabeled samples given a labeled dataset with known classe s. We exploit the peculiarities of NCD to build a new framework, named Neighborh ood Contrastive Learning (NCL), to learn discriminative representations that are important to clustering performance. Our contribution is twofold. First, we fin d that a feature extractor trained on the labeled set generates representations in which a generic query sample and its neighbors are likely to share the same c lass. We exploit this observation to retrieve and aggregate pseudo positive pair s with contrastive learning, thus encouraging the model to learn more discrimina tive representations. Second, we notice that most of the instances are easily di scriminated by the network, contributing less to the contrastive loss. To overco me this issue, we propose to generate hard negatives by mixing labeled and unlab eled samples in the feature space. We experimentally demonstrate that these two ingredients significantly contribute to clustering performance and lead our mode l to outperform state of the art by a large margin (e.g., clustering accuracy +1 3% on CIFAR-100 and +8% on ImageNet).
****
SimPoE: Simulated Character Control for 3D Human Pose Estimation
Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, Jason Saragih; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 7159-7169
Accurate estimation of 3D human motion from monocular video requires modeling bo th kinematics (body motion without physical forces) and dynamics (motion with ph ysical forces). To demonstrate this, we present SimPoE, a Simulation-based appro ach for 3D human Pose Estimation, which integrates image-based kinematic inferen ce and physics-based dynamics modeling. SimPoE learns a policy that takes as inp ut the current-frame pose estimate and the next image frame to control a physica lly-simulated character to output the next-frame pose estimate. The policy conta ins a learnable kinematic pose refinement unit that uses 2D keypoints to iterati vely refine its kinematic pose estimate of the next frame. Based on this refined kinematic pose, the policy learns to compute dynamics-based control (e.g., join t torques) of the character to advance the current-frame pose estimate to the po se estimate of the next frame. This design couples the kinematic pose refinement unit with the dynamics-based control generation unit, which are learned jointly with reinforcement learning to achieve accurate and physically-plausible pose e stimation. Furthermore, we propose a meta-control mechanism that dynamically adj usts the character's dynamics parameters based on the character state to attain more accurate pose estimates. Experiments on large-scale motion datasets demonst rate that our approach establishes the new state of the art in pose accuracy whi le ensuring physical plausibility.
****
Neural Camera Simulators
Hao Ouyang, Zifan Shi, Chenyang Lei, Ka Lung Law, Qifeng Chen; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 7700-7709
We present a controllable camera simulator based on deep neural networks to synt hesize raw image data under different camera settings, including exposure time, ISO, and aperture. The proposed simulator includes an exposure module that utili

zes the principle of modern lens designs for correcting the luminance level. It also contains a noise module using the noise level function and an aperture module with adaptive attention to simulate the side effects on noise and defocus blur. To facilitate the learning of a simulator model, we collect a dataset of the 10,000 raw images of 450 scenes with different exposure settings. Quantitative experiments and qualitative comparisons show that our approach outperforms relevant baselines in raw data synthesize on multiple cameras. Furthermore, the camera simulator enables various applications, including large-aperture enhancement, HDR, auto exposure, and data augmentation for training local feature detectors. Our work represents the first attempt to simulate a camera sensor's behavior leveraging both the advantage of traditional raw sensor features and the power of data-driven deep learning.
****
Neighborhood Normalization for Robust Geometric Feature Learning
Xingtong Liu, Benjamin D. Killeen, Ayushi Sinha, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, Mathias Unberath; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13049-13058
Extracting geometric features from 3D models is a common first step in applications such as 3D registration, tracking, and scene flow estimation. Many hand-crafted and learning-based methods aim to produce consistent and distinguishable geometric features for 3D models with partial overlap. These methods work well in cases where the point density and scale of the overlapping 3D objects are similar, but struggle in applications where 3D data are obtained independently with unknown global scale and scene overlap. Unfortunately, instances of this resolution mismatch are common in practice, e.g., when aligning data from multiple sensors. In this work, we introduce a new normalization technique, Batch-Neighborhood Normalization, aiming to improve robustness to mean-std variation of local feature distributions that presumably can happen in samples with varying point density. We empirically demonstrate that the presented normalization method's performance compares favorably to comparison methods in indoor and outdoor environments, and on a clinical dataset, on common point registration benchmarks in both standard and, particularly, resolution-mismatch settings. The source code and clinical dataset are available at https://github.com/lppllppl920/NeighborhoodNormalization-Pytorch.
****
Video Rescaling Networks With Joint Optimization Strategies for Downscaling and Upscaling
Yan-Cheng Huang, Yi-Hsin Chen, Cheng-You Lu, Hui-Po Wang, Wen-Hsiao Peng, Ching-Chun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3527-3536
This paper addresses the video rescaling task, which arises from the needs of adapting the video spatial resolution to suit individual viewing devices. We aim to jointly optimize video downscaling and upscaling as a combined task. Most recent studies focus on image-based solutions, which do not consider temporal information. We present two joint optimization approaches based on invertible neural networks with coupling layers. Our Long Short-Term Memory Video Rescaling Network (LSTM-VRN) leverages temporal information in the low-resolution video to form an explicit prediction of the missing high-frequency information for upscaling. Our Multi-input Multi-output Video Rescaling Network (MIMO-VRN) proposes a new strategy for downscaling and upscaling a group of video frames simultaneously. Not only do they outperform the image-based invertible model in terms of quantitative and qualitative results, but also show much improved upscaling quality than the video rescaling methods without joint optimization. To our best knowledge, this work is the first attempt at the joint optimization of video downscaling and upscaling.
****
TPCN: Temporal Point Cloud Networks for Motion Forecasting
Maosheng Ye, Tongyi Cao, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11318-11327
We propose the Temporal Point Cloud Networks (TPCN), a novel and flexible framew

ork with joint spatial and temporal learning for trajectory prediction. Unlike existing approaches that rasterize agents and map information as 2D images or operate in a graph representation, our approach extends ideas from point cloud learning with dynamic temporal learning to capture both spatial and temporal information by splitting trajectory prediction into both spatial and temporal dimensions. In the spatial dimension, agents can be viewed as an unordered point set, and thus it is straightforward to apply point cloud learning techniques to model agents' locations. While the spatial dimension does not take kinematic and motion information into account, we further propose dynamic temporal learning to model agents' motion over time. Experiments on the Argoverse motion forecasting benchmark show that our approach achieves state-of-the-art results.
****
TSGCNet: Discriminative Geometric Feature Learning With Two-Stream Graph Convolutional Network for 3D Dental Model Segmentation
Lingming Zhang, Yue Zhao, Deyu Meng, Zhiming Cui, Chenqiang Gao, Xinbo Gao, Chunfeng Lian, Dinggang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6699-6708
The ability to segment teeth precisely from digitized 3D dental models is an essential task in computer-aided orthodontic surgical planning. To date, deep learning based methods have been popularly used to handle this task. State-of-the-art methods directly concatenate the raw attributes of 3D inputs, namely coordinates and normal vectors of mesh cells, to train a single-stream network for fully-automated tooth segmentation. This, however, has the drawback of ignoring the different geometric meanings provided by those raw attributes. This issue might possibly confuse the network in learning discriminative geometric features and result in many isolated false predictions on the dental model. Against this issue, we propose a two-stream graph convolutional network (TSGCNet) to learn multi-view geometric information from different geometric attributes. Our TSGCNet adopts two graph-learning streams, designed in an input-aware fashion, to extract more discriminative high-level geometric representations from coordinates and normal vectors, respectively. These feature representations learned from the designed two different streams are further fused to integrate the multi-view complementary information for the cell-wise dense prediction task. We evaluate our proposed TSGCNet on a real-patient dataset of dental models acquired by 3D intraoral scanners, and experimental results demonstrate that our method significantly outperforms state-of-the-art methods for 3D shape segmentation.
****
Meta Batch-Instance Normalization for Generalizable Person Re-Identification
Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, Changick Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3425-3435
Although supervised person re-identification (Re-ID) methods have shown impressive performance, they suffer from a poor generalization capability on unseen domains. Therefore, generalizable Re-ID has recently attracted growing attention. Many existing methods have employed an instance normalization technique to reduce style variations, but the loss of discriminative information could not be avoided. In this paper, we propose a novel generalizable Re-ID framework, named Meta Batch-Instance Normalization (MetaBIN). Our main idea is to generalize normalization layers by simulating unsuccessful generalization scenarios beforehand in the meta-learning pipeline. To this end, we combine learnable batch-instance normalization layers with meta-learning and investigate the challenging cases caused by both batch and instance normalization layers. Moreover, we diversify the virtual simulations via our meta-train loss accompanied by a cyclic inner-updating manner to boost generalization capability. After all, the MetaBIN framework prevents our model from overfitting to the given source styles and improves the generalization capability to unseen domains without additional data augmentation or complicated network design. Extensive experimental results show that our model outperforms the state-of-the-art methods on the large-scale domain generalization Re-ID benchmark and the cross-domain Re-ID problem. The source code is available at: https://github.com/bismex/MetaBIN.

****
Dictionary-Guided Scene Text Recognition
Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7383-7392
Language prior plays an important role in the way humans perceive and recognize text in the wild. In this work, we present an approach to train and use scene text recognition models by exploiting multiple clues from a language reference. Current scene text recognition methods have used lexicons to improve recognition performance, but their naive approach of simply casting the output into a dictionary word based purely on the edit distance has many limitations. We introduce here a novel approach to incorporate a dictionary in both the training and inference stage of a scene text recognition system. We use the dictionary to generate a list of possible outcomes and find the one that is most compatible with the visual appearance of the text. The proposed method leads to a robust scene text recognition model, which is better at handling ambiguous cases encountered in the wild, and improves the overall performance of a state-of-the-art scene text spotting framework. Our work suggests that incorporating language prior is a potential approach to advance scene text detection and recognition methods. Besides, we contribute a challenging scene text dataset for Vietnamese, where some characters are equivocal in the visual form due to accent symbols. This dataset will serve as a challenging benchmark for measuring the applicability and robustness of scene text detection and recognition algorithms.
****
Glance and Gaze: Inferring Action-Aware Points for One-Stage Human-Object Interaction Detection
Xubin Zhong, Xian Qu, Changxing Ding, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13234-13243
Modern human-object interaction (HOI) detection approaches can be divided into one-stage methods and two-stage ones. One-stage models are more efficient due to their straightforward architectures, but the two-stage models are still advantageous in accuracy. Existing one-stage models usually begin by detecting predefined interaction areas or points, and then attend to these areas only for interaction prediction; therefore, they lack reasoning steps that dynamically search for discriminative cues. In this paper, we propose a novel one-stage method, namely Glance and Gaze Network (GGNet), which adaptively models a set of action-aware points (ActPoints) via glance and gaze steps. The glance step quickly determines whether each pixel in the feature maps is an interaction point. The gaze step leverages feature maps produced by the glance step to adaptively infer ActPoints around each pixel in a progressive manner. Features of the refined ActPoints are aggregated for interaction prediction. Moreover, we design an action-aware approach that effectively matches each detected interaction with its associated human-object pair, along with a novel hard negative attentive loss to improve the optimization of GGNet. All the above operations are conducted simultaneously and efficiently for all pixels in the feature maps. Finally, GGNet outperforms state-of-the-art methods by significant margins on both V-COCO and HICO-DET benchmarks. Code of GGNet is available at https://github.com/SherlockHolmes221/GGNet.
****
Activate or Not: Learning Customized Activation
Ningning Ma, Xiangyu Zhang, Ming Liu, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8032-8042
We present a simple, effective, and general activation function we term ACON which learns to activate the neurons or not. Interestingly, we find Swish, the recent popular NAS-searched activation, can be interpreted as a smooth approximation to ReLU. Intuitively, in the same way, we approximate the more general Maxout family to our novel ACON family, which remarkably improves the performance and makes Swish a special case of ACON. Next, we present meta-ACON, which explicitly learns to optimize the parameter switching between non-linear (activate) and linear (inactivate) and provides a new design space. By simply changing the activati

on function, we show its effectiveness on both small models and highly optimized large models (e.g. it improves the ImageNet top-1 accuracy rate by 6.7% and 1.8% on MobileNet-0.25 and ResNet-152, respectively). Moreover, our novel ACON can be naturally transferred to object detection and semantic segmentation, showing that ACON is an effective alternative in a variety of tasks. Code is available at https://github.com/nmaac/acon.
****
Wide-Baseline Relative Camera Pose Estimation With Directional Learning
Kefan Chen, Noah Snavely, Ameesh Makadia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3258-3268
Modern deep learning techniques that regress the relative camera pose between two images have difficulty dealing with challenging scenarios, such as large camera motions resulting in occlusions and significant changes in perspective that leave little overlap between images. These models continue to struggle even with the benefit of large supervised training datasets. To address the limitations of these models, we take inspiration from techniques that show regressing keypoint locations in 2D and 3D can be improved by estimating a discrete distribution over keypoint locations. Analogously, in this paper we explore improving camera pose regression by instead predicting a discrete distribution over camera poses. To realize this idea, we introduce DirectionNet, which estimates discrete distributions over the 5D relative pose space using a novel parameterization to make the estimation problem tractable. Specifically, DirectionNet factorizes relative camera pose, specified by a 3D rotation and a translation direction, into a set of 3D direction vectors. Since 3D directions can be identified with points on the sphere, DirectionNet estimates discrete distributions on the sphere as its output. We evaluate our model on challenging synthetic and real pose estimation datasets constructed from Matterport3D and InteriorNet. Promising results show a near 50% reduction in error over direct regression methods.
****
Improving Unsupervised Image Clustering With Robust Learning
Sungwon Park, Sungwon Han, Sundong Kim, Danu Kim, Sungkyu Park, Seunghoon Hong, Meeyoung Cha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12278-12287
Unsupervised image clustering methods often introduce alternative objectives to indirectly train the model and are subject to faulty predictions and overconfident results. To overcome these challenges, the current research proposes an innovative model RUC that is inspired by robust learning. RUC's novelty is at utilizing pseudo-labels of existing image clustering models as a noisy dataset that may include misclassified samples. Its retraining process can revise misaligned knowledge and alleviate the overconfidence problem in predictions. The model's flexible structure makes it possible to be used as an add-on module to other clustering methods and helps them achieve better performance on multiple datasets. Extensive experiments show that the proposed model can adjust the model confidence with better calibration and gain additional robustness against adversarial noise.
****
Neural Surface Maps
Luca Morreale, Noam Aigerman, Vladimir G. Kim, Niloy J. Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4639-4648
Maps are arguably one of the most fundamental concepts used to define and operate on manifold surfaces in differentiable geometry. Accordingly, in geometry processing, maps are ubiquitous and are used in many core applications, such as paramterization, shape analysis, remeshing, and deformation. Unfortunately, most computational representations of surface maps do not lend themselves to manipulation and optimization, usually entailing hard, discrete problems. While algorithms exist to solve these problems, they are problem-specific, and a general framework for surface maps is still in need. In this paper, we advocate to consider neural networks as encoding surface maps. Since neural networks can be composed on one another and are differentiable, we show it is easy to use them to define surfaces via atlases, compose them for surface-to-surface mappings, and optimize dif

ferentiable objectives relating to them, such as any notion of distortion, in a trivial manner. In our experiments, we represent surfaces by generating a neural map that approximates a UV parameterization of a 3D model. Then, we compose this map with other neural maps which we optimize with respect to distortion measures. We show that our formulation enables trivial optimization of rather elusive mapping tasks, such as maps between a collection of surfaces.
****
Enhance Curvature Information by Structured Stochastic Quasi-Newton Methods
Minghan Yang, Dong Xu, Hongyu Chen, Zaiwen Wen, Mengyun Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10654-10663
In this paper, we consider stochastic second-order methods for minimizing a finite summation of nonconvex functions. One important key is to find an ingenious but cheap scheme to incorporate local curvature information. Since the true Hessian matrix is often a combination of a cheap part and an expensive part, we propose a structured stochastic quasi-Newton method by using partial Hessian information as much as possible. By further exploiting either the low-rank structure or the Kronecker-product properties of the quasi-Newton approximations, the computation of the quasi-Newton direction is affordable. Global convergence to stationary point and local superlinear convergence rate are established under some mild assumptions. Numerical results on logistic regression, deep autoencoder networks and deep convolutional neural networks show that our proposed method is quite competitive to the state-of-the-art methods.
****
Variational Relational Point Completion Network
Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8524-8533
Real-scanned point clouds are often incomplete due to viewpoint, occlusion, and noise. Existing point cloud completion methods tend to generate global shape skeletons and hence lack fine local details. Furthermore, they mostly learn a deterministic partial-to-complete mapping, but overlook structural relations in man-made objects. To tackle these challenges, this paper proposes a variational framework, Variational Relational point Completion network (VRCNet) with two appealing properties: 1) Probabilistic Modeling. In particular, we propose a dual-path architecture to enable principled probabilistic modeling across partial and complete clouds. One path consumes complete point clouds for reconstruction by learning a point VAE. The other path generates complete shapes for partial point clouds, whose embedded distribution is guided by distribution obtained from the reconstruction path during training. 2) Relational Enhancement. Specifically, we carefully design point self-attention kernel and point selective kernel module to exploit relational point features, which refines local shape details conditioned on the coarse completion. In addition, we contribute a multi-view partial point cloud dataset (MVP dataset) containing over 100,000 high-quality scans, which renders partial 3D shapes from 26 uniformly distributed camera poses for each 3D CAD model. Extensive experiments demonstrate that VRCNet outperforms state-of-the-art methods on all standard point cloud completion benchmarks. Notably, VRCNet shows great generalizability and robustness on real-world point cloud scans.
****
StruMonoNet: Structure-Aware Monocular 3D Prediction
Zhenpei Yang, Li Erran Li, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7413-7422
Monocular 3D prediction is one of the fundamental problems in 3D vision. Recent deep learning-based approaches have brought us exciting progress on this problem. However, existing approaches have predominantly focused on end-to-end depth and normal predictions, which do not fully utilize the underlying 3D environment's geometric structures. This paper introduces StruMonoNet, which detects and enforces a planar structure to enhance pixel-wise predictions. StruMonoNet innovates in leveraging a hybrid representation that combines visual feature and a surfel representation for plane prediction. This formulation allows us to combine the

power of visual feature learning and the flexibility of geometric representation s in incorporating geometric relations. As a result, StruMonoNet can detect relations between planes such as adjacent planes, perpendicular planes, and parallel planes, all of which are beneficial for dense 3D prediction. Experimental results show that StruMonoNet considerably outperforms state-of-the-art approaches on NYUv2 and ScanNet.
****
Learning To Relate Depth and Semantics for Unsupervised Domain Adaptation
Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8197-8207
We present an approach for encoding visual task relationships to improve model performance in an Unsupervised Domain Adaptation (UDA) setting. Semantic segmentation and monocular depth estimation are shown to be complementary tasks; in a multi-task learning setting, a proper encoding of their relationships can further improve performance on both tasks. Motivated by this observation, we propose a novel Cross-Task Relation Layer (CTRL), which encodes task dependencies between the semantic and depth predictions. To capture the cross-task relationships, we propose a neural network architecture that contains task-specific and cross-task refinement heads. Furthermore, we propose an Iterative Self-Learning (ISL) training scheme, which exploits semantic pseudo-labels to provide extra supervision on the target domain. We experimentally observe improvements in both tasks' performance because the complementary information present in these tasks is better captured. Specifically, we show that: (1) our approach improves performance on all tasks when they are complementary and mutually dependent; (2) the CTRL helps to improve both semantic segmentation and depth estimation tasks performance in the challenging UDA setting; (3) the proposed ISL training scheme further improves the semantic segmentation performance. The implementation is available at https://github.com/susaha/ctrl-uda.
****
Training Networks in Null Space of Feature Covariance for Continual Learning
Shipeng Wang, Xiaorong Li, Jian Sun, Zongben Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 184-193
In the setting of continual learning, a network is trained on a sequence of tasks, and suffers from catastrophic forgetting. To balance plasticity and stability of network in continual learning, in this paper, we propose a novel network training algorithm called Adam-NSCL, which sequentially optimizes network parameters in the null space of previous tasks. We first propose two mathematical conditions respectively for achieving network stability and plasticity in continual learning. Based on them, the network training for sequential tasks can be simply achieved by projecting the candidate parameter update into the approximate null space of all previous tasks in the network training process, where the candidate parameter update can be generated by Adam. The approximate null space can be derived by applying singular value decomposition to the uncentered covariance matrix of all input features of previous tasks for each linear layer. For efficiency, the uncentered covariance matrix can be incrementally computed after learning each task. We also empirically verify the rationality of the approximate null space at each linear layer. We apply our approach to training networks for continual learning on benchmark datasets of CIFAR-100 and TinyImageNet, and the results suggest that the proposed approach outperforms or matches the state-ot-the-art continual learning approaches.
****
PiCIE: Unsupervised Semantic Segmentation Using Invariance and Equivariance in Clustering
Jang Hyun Cho, Utkarsh Mall, Kavita Bala, Bharath Hariharan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16794-16804
We present a new framework for semantic segmentation without annotations via clustering. Off-the-shelf clustering methods are limited to curated, single-label, and object-centric images yet real-world data are dominantly uncurated, multi-la

bel, and scene-centric. We extend clustering from images to pixels and assign se
parate cluster membership to different instances within each image. However, sol
ely relying on pixel-wise feature similarity fails to learn high-level semantic
concepts and overfits to low-level visual cues. We propose a method to incorpora
te geometric consistency as an inductive bias to learn invariance and equivarian
ce for photometric and geometric variations. With our novel learning objective,
our framework can learn high-level semantic concepts. Our method, PiCIE (Pixel-l
evel feature Clustering using Invariance and Equivariance), is the first method
capable of segmenting both things and stuff categories without any hyperparamete
r tuning or task-specific pre-processing. Our method largely outperforms existin
g baselines on COCO and Cityscapes with +17.5 Acc. and +4.5 mIoU. We show that P
iCIE gives a better initialization for standard supervised training. The code is
 available at https:// github.com/janghyuncho/PiCIE.
****
DyCo3D: Robust Instance Segmentation of 3D Point Clouds Through Dynamic Convolut
ion
Tong He, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 354-363
Previous top-performing approaches for point cloud instance segmentation involve
 a bottom-up strategy, which often includes inefficient operations or complex pi
pelines, such as grouping over-segmented components, introducing additional step
s for refining, or designing complicated loss functions. The inevitable variatio
n in the instance scales can lead bottom-up methods to become particularly sensi
tive to hyper-parameter values. To this end, we propose instead a dynamic, propo
sal-free, data-driven approach that generates the appropriate convolution kernel
s to apply in response to the nature of the instances. To make the kernels discr
iminative, we explore a large context by gathering homogeneous points that share
 identical semantic categories and have close votes for the geometric centroids.
 Instances are then decoded by several simple convolutional layers. Due to the l
imited receptive field introduced by the sparse convolution, a small light-weigh
t transformer is also devised to capture the long-range dependencies and high-le
vel interactions among point samples. The proposed method achieves promising res
ults on both ScanetNetV2 and S3DIS, and this performance is robust to the partic
ular hyper-parameter values chosen. It also improves inference speed by more tha
n 25% over the current state-of-the-art. Code is available at: https://git.io/Dy
Co3D
****
SSLayout360: Semi-Supervised Indoor Layout Estimation From 360deg Panorama
Phi Vu Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2021, pp. 15353-15362
Recent years have seen flourishing research on both semi-supervised learning and
 3D room layout reconstruction. In this work, we explore the intersection of the
se two fields to advance the research objective of enabling more accurate 3D ind
oor scene modeling with less labeled data. We propose the first approach to lear
n representations of room corners and boundaries by using a combination of label
ed and unlabeled data for improved layout estimation in a 360-degree panoramic s
cene. Through extensive comparative experiments, we demonstrate that our approac
h can advance layout estimation of complex indoor scenes using as few as 20 labe
led examples. When coupled with a layout predictor pre-trained on synthetic data
, our semi-supervised method matches the fully supervised counterpart using only
 12% of the labels. Our work takes an important first step towards robust semi-s
upervised layout estimation that can enable many applications in 3D perception w
ith limited labeled data.
****
SLADE: A Self-Training Framework for Distance Metric Learning
Jiali Duan, Yen-Liang Lin, Son Tran, Larry S. Davis, C.-C. Jay Kuo; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
21, pp. 9644-9653
Most existing distance metric learning approaches use fully labeled data to lear
n the sample similarities in an embedding space. We present a self-training fram

ework, SLADE, to improve retrieval performance by leveraging additional unlabeled data. We first train a teacher model on the labeled data and use it to generate pseudo labels for the unlabeled data. We then train a student model on both labels and pseudo labels to generate final feature embeddings. We use self-supervised representation learning to initialize the teacher model. To better deal with noisy pseudo labels generated by the teacher network, we design a new feature basis learning component for the student network, which learns basis functions of feature representations for unlabeled data. The learned basis vectors better measure the pairwise similarity and are used to select high-confident samples for training the student network. We evaluate our method on standard retrieval benchmarks: CUB-200, Cars-196 and In-shop. Experimental results demonstrate that with additional unlabeled data, our approach significantly improves the performance over the state-of-the-art methods.
****
NormalFusion: Real-Time Acquisition of Surface Normals for High-Resolution RGB-D Scanning
Hyunho Ha, Joo Ho Lee, Andreas Meuleman, Min H. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15970-15979
Multiview shape-from-shading (SfS) has achieved high-detail geometry, but its computation is expensive for solving a multiview registration and an ill-posed inverse rendering problem. Therefore, it has been mainly used for offline methods. Volumetric fusion enables real-time scanning using a conventional RGB-D camera, but its geometry resolution has been limited by the grid resolution of the volumetric distance field and depth registration errors. In this paper, we propose a real-time scanning method that can acquire high-detail geometry by bridging volumetric fusion and multiview SfS in two steps. First, we propose the first real-time acquisition of photometric normals stored in texture space to achieve high-detail geometry. We also introduce geometry-aware texture mapping, which progressively refines geometric registration between the texture space and the volumetric distance field by means of normal texture, achieving real-time multiview SfS. We demonstrate our scanning of high-detail geometry using an RGB-D camera at 20 fps. Results verify that the geometry quality of our method is strongly competitive with that of offline multi-view SfS methods.
****
SE-SSD: Self-Ensembling Single-Stage Object Detector From Point Cloud
Wu Zheng, Weiliang Tang, Li Jiang, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14494-14503
We present Self-Ensembling Single-Stage object Detector (SE-SSD) for accurate and efficient 3D object detection in outdoor point clouds. Our key focus is on exploiting both soft and hard targets with our formulated constraints to jointly optimize the model, without introducing extra computation in the inference. Specifically, SE-SSD contains a pair of teacher and student SSDs, in which we design an effective IoU-based matching strategy to filter soft targets from the teacher and formulate a consistency loss to align student predictions with them. Also, to maximize the distilled knowledge for ensembling the teacher, we design a new augmentation scheme to produce shape-aware augmented samples to train the student, aiming to encourage it to infer complete object shapes. Lastly, to better exploit hard targets, we design an ODIoU loss to supervise the student with constraints on the predicted box centers and orientations. Our SE-SSD attains top performance compared with all prior published works. Also, it attains top precisions for car detection in the KITTI benchmark (ranked 1st and 2nd on the BEV and 3D leaderboards, respectively) with an ultra-high inference speed. The code is available at https://github.com/Vegeta2020/SE-SSD.
****
Where and What? Examining Interpretable Disentangled Representations
Xinqi Zhu, Chang Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5861-5870
Capturing interpretable variations has long been one of the goals in disentanglement learning. However, unlike the independence assumption, interpretability has

rarely been exploited to encourage disentanglement in the unsupervised setting. In this paper, we examine the interpretability of disentangled representations by investigating two questions: where to be interpreted and what to be interpreted? A latent code is easily to be interpreted if it would consistently impact a certain subarea of the resulting generated image. We thus propose to learn a spatial mask to localize the effect of each individual latent dimension. On the other hand, interpretability usually comes from latent dimensions that capture simple and basic variations in data. We thus impose a perturbation on a certain dimension of the latent code, and expect to identify the perturbation along this dimension from the generated images so that the encoding of simple variations can be enforced. Additionally, we develop an unsupervised model selection method, which accumulates perceptual distance scores along axes in the latent space. On various datasets, our models can learn high-quality disentangled representations without supervision, showing the proposed modeling of interpretability is an effective proxy for achieving unsupervised disentanglement.
****
Physically-Aware Generative Network for 3D Shape Modeling
Mariem Mezghanni, Malika Boulkenafed, Andre Lieutier, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9330-9341
Shapes are often designed to satisfy structural properties and serve a particular functionality in the physical world. Unfortunately, most existing generative models focus primarily on the geometric or visual plausibility, ignoring the physical or structural constraints. To remedy this, we present a novel method aimed to endow deep generative models with physical reasoning. In particular, we introduce a loss and a learning framework that promote two key characteristics of the generated shapes: their connectivity and physical stability. The former ensures that each generated shape consists of a single connected component, while the latter promotes the stability of that shape when subjected to gravity. Our proposed physical losses are fully differentiable and we demonstrate their use in end-to-end learning. Crucially we demonstrate that such physical objectives can be achieved without sacrificing the expressive power of the model and variability of the generated results. We demonstrate through extensive comparisons with the state-of-the-art deep generative models, the utility and efficiency of our proposed approach, while avoiding the potentially costly differentiable physical simulation at training time.
****
Bilinear Parameterization for Non-Separable Singular Value Penalties
Marcus Valtonen Ornhag, Jose Pedro Iglesias, Carl Olsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3897-3906
Low rank inducing penalties have been proven to successfully uncover fundamental structures considered in computer vision and machine learning; however, such methods generally lead to non-convex optimization problems. Since the resulting objective is non-convex one often resorts to using standard splitting schemes such as Alternating Direction Methods of Multipliers (ADMM), or other subgradient methods, which exhibit slow convergence in the neighbourhood of a local minimum. We propose a method using second order methods, in particular the variable Projection method (VarPro), by replacing the non-convex penalties with a surrogate capable of converting the original objectives to differentiable equivalents. In this way we benefit from faster convergence. The bilinear framework is compatible with a large family of regularizers, and we demonstrate the benefits of our approach on real datasets for rigid and non-rigid structure from motion. The qualitative difference in reconstructions show that many popular non-convex objectives enjoy an advantage in transitioning to the proposed framework.
****
Objectron: A Large Scale Dataset of Object-Centric Videos in the Wild With Pose Annotations
Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, Matthias Grundmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn

ition (CVPR), 2021, pp. 7822-7831
3D object detection has recently become popular due to many applications in robo tics, augmented reality, autonomy, and image retrieval. We introduce the Objectr on dataset to advance the state of the art in 3D object detection and foster new research and applications, such as 3D object tracking, view synthesis, and impr oved 3D shape representation. The dataset contains object-centric short videos w ith pose annotations for nine categories and includes 4 million annotated images in 14,819 annotated videos. We also propose a new evaluation metric, 3D Interse ction over Union, for 3D object detection. We demonstrate the usefulness of our dataset in 3D object detection and novel view synthesis tasks by providing basel ine models trained on this dataset. Our dataset and evaluation source code are a vailable online at Github.com/google-research-datasets/Objectron.
****
Intra-Inter Camera Similarity for Unsupervised Person Re-Identification
Shiyu Xuan, Shiliang Zhang; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2021, pp. 11926-11935
Most of unsupervised person Re-Identification (Re-ID) works produce pseudo-label s by measuring the feature similarity without considering the distribution discr epancy among cameras, leading to degraded accuracy in label computation across c ameras. This paper targets to address this challenge by studying a novel intra-i nter camera similarity for pseudo-label generation. We decompose the sample simi larity computation into two stage, i.e., the intra-camera and inter-camera compu tations, respectively. The intra-camera computation directly leverages the CNN f eatures for similarity computation within each camera. Pseudo-labels generated o n different cameras train the re-id model in a multi-branch network. The second stage considers the classification scores of each sample on different cameras as a new feature vector. This new feature effectively alleviates the distribution discrepancy among cameras and generates more reliable pseudo-labels. We hence tr ain our re-id model in two stages with intra-camera and inter-camera pseudo-labe ls, respectively. This simple intra-inter camera similarity produces surprisingl y good performance on multiple datasets, e.g., achieves rank-1 accuracy of 89.5% on the Market1501 dataset, outperforming the recent unsupervised works by 9+%, and is comparable with the latest transfer learning works that leverage extra an notations.
****
Efficient Feature Transformations for Discriminative and Generative Continual Le arning
Vinay Kumar Verma, Kevin J Liang, Nikhil Mehta, Piyush Rai, Lawrence Carin; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2021, pp. 13865-13875
As neural networks are increasingly being applied to real-world applications, me chanisms to address distributional shift and sequential task learning without fo rgetting are critical. Methods incorporating network expansion have shown promis e by naturally adding model capacity for learning new tasks while simultaneously avoiding catastrophic forgetting. However, the growth in the number of addition al parameters of many of these types of methods can be computationally expensive at larger scales, at times prohibitively so. Instead, we propose a simple task- specific feature map transformation strategy for continual learning, which we ca ll Efficient Feature Transformations (EFTs). These EFTs provide powerful flexibi lity for learning new tasks, achieved with minimal parameters added to the base architecture. We further propose a feature distance maximization strategy, which significantly improves task prediction in class incremental settings, without n eeding expensive generative models. We demonstrate the efficacy and efficiency o f our method with an extensive set of experiments in discriminative (CIFAR-100 a nd ImageNet-1K) and generative (LSUN, CUB-200, Cats) sequences of tasks. Even wi th low single-digit parameter growth rates, EFTs can outperform many other conti nual learning methods in a wide range of settings.
****
Learning a Self-Expressive Network for Subspace Clustering
Shangzhi Zhang, Chong You, Rene Vidal, Chun-Guang Li; Proceedings of the IEEE/CV

F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12393-12403
State-of-the-art subspace clustering methods are based on the self-expressive model, which represents each data point as a linear combination of other data points. However, such methods are designed for a finite sample dataset and lack the ability to generalize to out-of-sample data. Moreover, since the number of self-expressive coefficients grows quadratically with the number of data points, their ability to handle large-scale datasets is often limited. In this paper, we propose a novel framework for subspace clustering, termed Self-Expressive Network (SENet), which employs a properly designed neural network to learn a self-expressive representation of the data. We show that our SENet can not only learn the self-expressive coefficients with desired properties on the training data, but also handle out-of-sample data. Besides, we show that SENet can also be leveraged to perform subspace clustering on large-scale datasets. Extensive experiments conducted on synthetic data and real world benchmark data validate the effectiveness of the proposed method. In particular, SENet yields highly competitive performance on MNIST, Fashion MNIST and Extended MNIST and state-of-the-art performance on CIFAR-10.
****
A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning
Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, Kaiming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3299-3309
We present a large-scale study on unsupervised spatiotemporal representation learning from videos. With a unified perspective on four recent image-based frameworks, we study a simple objective that can easily generalize all these methods to space-time. Our objective encourages temporally-persistent features in the same video, and in spite of its simplicity, it works surprisingly well across: (i) different unsupervised frameworks, (ii) pre-training datasets, (iii) downstream datasets, and (iv) backbone architectures. We draw a series of intriguing observations from this study, e.g., we discover that encouraging long-spanned persistency can be effective even if the timespan is 60 seconds. In addition to state-of-the-art results in multiple benchmarks, we report a few promising cases in which unsupervised pre-training can outperform its supervised counterpart. Code will be made available at https://github.com/facebookresearch/SlowFast.
****
Asymmetric Metric Learning for Knowledge Transfer
Mateusz Budnik, Yannis Avrithis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8228-8238
Knowledge transfer from large teacher models to smaller student models has recently been studied for metric learning, focusing on fine-grained classification. In this work, focusing on instance-level image retrieval, we study an asymmetric testing task, where the database is represented by the teacher and queries by the student. Inspired by this task, we introduce asymmetric metric learning, a novel paradigm of using asymmetric representations at training. This acts as a simple combination of knowledge transfer with the original metric learning task. We systematically evaluate different teacher and student models, metric learning and knowledge transfer loss functions on the new asymmetric testing as well as the standard symmetric testing task, where database and queries are represented by the same model. We find that plain regression is surprisingly effective compared to more complex knowledge transfer mechanisms, working best in asymmetric testing. Interestingly, our asymmetric metric learning approach works best in symmetric testing, allowing the student to even outperform the teacher. Our implementation is publicly available, including trained student models for all loss functions and all pairs of teacher/student models. This can serve as a benchmark for future research.
****
Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection
Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, Yongdong Zhang; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6458-6467
Face forgery detection is raising ever-increasing interest in computer vision since facial manipulation technologies cause serious worries. Though recent works have reached sound achievements, there are still unignorable problems: a) learned features supervised by softmax loss are separable but not discriminative enough, since softmax loss does not explicitly encourage intra-class compactness and interclass separability; and b) fixed filter banks and hand-crafted features are insufficient to capture forgery patterns of frequency from diverse inputs. To compensate for such limitations, a novel frequency-aware discriminative feature learning framework is proposed in this paper. Specifically, we design a novel single-center loss (SCL) that only compresses intra-class variations of natural faces while boosting interclass differences in the embedding space. In such a case, the network can learn more discriminative features with less optimization difficulty. Besides, an adaptive frequency feature generation module is developed to mine frequency clues in a completely data-driven fashion. With the above two modules, the whole framework can learn more discriminative features in an end-to-end manner. Extensive experiments demonstrate the effectiveness and superiority of our framework on three versions of the FF++ dataset.
****
3DCaricShop: A Dataset and a Baseline Method for Single-View 3D Caricature Face Reconstruction
Yuda Qiu, Xiaojie Xu, Lingteng Qiu, Yan Pan, Yushuang Wu, Weikai Chen, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10236-10245
Caricature is an artistic representation that deliberately exaggerates the distinctive features of a human face to convey humor or sarcasm. However, reconstructing a 3D caricature from a 2D caricature image remains a challenging task, mostly due to the lack of data. We propose to fill this gap by introducing 3DCaricShop, the first large-scale 3D caricature dataset that contains 2000 high-quality diversified 3D caricatures manually crafted by professional artists. 3DCaricShop also provides rich annotations including a paired 2D caricature image, camera parameters, and 3D facial landmarks. To demonstrate the advantage of 3DCaricShop, we present a novel baseline approach for single-view 3D caricature reconstruction. To ensure a faithful reconstruction with plausible face deformations, we propose to connect the good ends of the detail-rich implicit functions and the parametric mesh representations. In particular, we first register a template mesh to the output of the implicit generator and iteratively project the registration result onto a pre-trained PCA space to resolve artifacts and self-intersections. To deal with the large deformation during non-rigid registration, we propose a novel view-collaborative graph convolution network (VC-GCN) to extract key points from the implicit mesh for accurate alignment. Our method is able to generate high-fidelity 3D caricature in a pre-defined mesh topology that is animation-ready. Extensive experiments have been conducted on 3DCaricShop to verify the significance of the database and the effectiveness of the proposed method. We will release 3DCaricShop upon publication.
****
OCONet: Image Extrapolation by Object Completion
Richard Strong Bowen, Huiwen Chang, Charles Herrmann, Piotr Teterwak, Ce Liu, Ramin Zabih; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2307-2317
Image extrapolation extends an input image beyond the originally-captured field of view. Existing methods struggle to extrapolate images with salient objects in the foreground or are limited to very specific objects such as humans, but tend to work well on indoor/outdoor scenes. We introduce OCONet (Object COmpletion Networks) to extrapolate foreground objects, with an object completion network conditioned on its class. OCONet uses an encoder-decoder architecture trained with adversarial loss to predict the object's texture as well as its extent, represented as a predicted signed-distance field. An independent step extends the background, and the object is composited on top using the predicted mask. Both qualit

ative and quantitative results show that we improve on state-of-the-art image ex trapolation results for challenging examples.
****
VisualVoice: Audio-Visual Speech Separation With Cross-Modal Consistency
Ruohan Gao, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15495-15505
We introduce a new approach for audio-visual speech separation. Given a video, t he goal is to extract the speech associated with a face in spite of simultaneous background sounds and/or other human speakers. Whereas existing methods focus o n learning the alignment between the speaker's lip movements and the sounds they generate, we propose to leverage the speaker's face appearance as an additional prior to isolate the corresponding vocal qualities they are likely to produce. Our approach jointly learns audio-visual speech separation and cross-modal speak er embeddings from unlabeled video. It yields state-of-the-art results on five b enchmark datasets for audio-visual speech separation and enhancement, and genera lizes well to challenging real-world videos of diverse scenarios. Our video resu lts and code: http://vision.cs.utexas.edu/projects/VisualVoice/.
****
Fair Attribute Classification Through Latent Space De-Biasing
Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 930 1-9310
Fairness in visual recognition is becoming a prominent and critical topic of dis cussion as recognition systems are deployed at scale in the real world. Models t rained from data in which target labels are correlated with protected attributes (e.g., gender, race) are known to learn and exploit those correlations. In this work, we introduce a method for training accurate target classifiers while miti gating biases that stem from these correlations. We use GANs to generate realist ic-looking images, and perturb these images in the underlying latent space to ge nerate training data that is balanced for each protected attribute. We augment t he original dataset with this generated data, and empirically demonstrate that t arget classifiers trained on the augmented dataset exhibit a number of both quan titative and qualitative benefits. We conduct a thorough evaluation across multi ple target labels and protected attributes in the CelebA dataset, and provide an in-depth analysis and comparison to existing literature in the space. Code can be found at https://github.com/princetonvisualai/gan-debiasing.
****
Correlated Input-Dependent Label Noise in Large-Scale Image Classification
Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, Jesse Berent; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1551-1560
Large scale image classification datasets often contain noisy labels. We take a principled probabilistic approach to modelling input-dependent, also known as he teroscedastic, label noise in these datasets. We place a multivariate Normal dis tributed latent variable on the final hidden layer of a neural network classifie r. The covariance matrix of this latent variable, models the aleatoric uncertain ty due to label noise. We demonstrate that the learned covariance structure capt ures known sources of label noise between semantically similar and co-occurring classes. Compared to standard neural network training and other baselines, we sh ow significantly improved accuracy on Imagenet ILSVRC 2012 79.3% (+ 2.6%), Image net-21k 47.0% (+ 1.1%) and JFT 64.7% (+ 1.6%). We set a new state-of-the-art res ult on WebVision 1.0 with 76.6% top-1 accuracy. These datasets range from over 1 M to over 300M training examples and from 1k classes to more than 21k classes. O ur method is simple to use, and we provide an implementation that is a drop-in r eplacement for the final fully-connected layer in a deep classifier.
****
Delving Into Localization Errors for Monocular 3D Object Detection
Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, Wanli Ouyan g; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2021, pp. 4721-4730

Estimating 3D bounding boxes from monocular images is an essential component in
autonomous driving, while accurate 3D object detection from this kind of data is
 very challenging. In this work, by intensive diagnosis experiments, we quantify
 the impact introduced by each sub-task and found the `localization error' is th
e vital factor in restricting monocular 3D detection. Besides, we also investiga
te the underlying reasons behind localization errors, analyze the issues they mi
ght bring, and propose three strategies. First, we revisit the misalignment betw
een the center of the 2D bounding box and the projected center of the 3D object,
 which is a vital factor leading to low localization accuracy. Second, we observ
e that accurately localizing distant objects with existing technologies is almos
t impossible, while those samples will mislead the learned network. To this end,
 we propose to remove such samples from the training set for improving the overa
ll performance of the detector. Lastly, we also propose a novel 3D IoU oriented
loss for the size estimation of the object, which is not affected by `localizati
on error'. We conduct extensive experiments on the KITTI dataset, where the prop
osed method achieves real-time detection and outperforms previous methods by a l
arge margin. The code will be made available at: https://github.com/xinzhuma/mon
odle.
****
Nearest Neighbor Matching for Deep Clustering
Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, Heng Huang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1369
3-13702
Deep clustering gradually becomes an important branch in unsupervised learning m
ethods. However, current approaches hardly take into consideration the semantic
sample relationships that existed in both local and global features. In addition
, since the deep features are updated on-the-fly, relying on these sample relati
onships may construct more semantically confident sample pairs, leading to infer
ior performance. To tackle this issue, we propose a method called Nearest Neighb
or Matching (NNM) to match samples with their nearest neighbors from both local
(batch) and global (overall) levels. Specifically, for the local level, we match
 the nearest neighbors based on batch embedded features, as for the global one,
we match neighbors from overall embedded features. To keep the clustering assign
ment consistent in both neighbors and classes, we frame consistent loss and clas
s contrastive loss for both local and global levels. Experimental results on thr
ee benchmark datasets demonstrate the superiority of our new model against state
-of-the-art methods. Particularly on the STL-10 dataset, our method can achieve
supervised performance. As for the CIFAR-100 dataset, our NNM leads 3.7% against
 the latest comparison method. Our code will be available at https://github.com/
ZhiyuanDang/NNM.
****
MOOD: Multi-Level Out-of-Distribution Detection
Ziqian Lin, Sreya Dutta Roy, Yixuan Li; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15313-15323
Out-of-distribution (OOD) detection is essential to prevent anomalous inputs fro
m causing a model to fail during deployment. While improved OOD detection method
s have emerged, they often rely on the final layer outputs and require a full fe
edforward pass for any given input. In this paper, we propose a novel framework,
 multi-level out-of-distribution detection MOOD, which exploits intermediate cla
ssifier outputs for dynamic and efficient OOD inference. We explore and establis
h a direct relationship between the OOD data complexity and optimal exit level,
and show that easy OOD examples can be effectively detected early without propag
ating to deeper layers. At each exit, the OOD examples can be distinguished thro
ugh our proposed adjusted energy score, which is both empirically and theoretica
lly suitable for networks with multiple classifiers. We extensively evaluate MOO
D across 10 OOD datasets spanning a wide range of complexities. Experiments demo
nstrate that MOOD achieves up to 71.05% computational reduction in inference, wh
ile maintaining competitive OOD detection performance.
****
Equalization Loss v2: A New Gradient Balance Approach for Long-Tailed Object Det

ection
Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, Quanquan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1685-1694
Recently proposed decoupled training methods emerge as a dominant paradigm for long-tailed object detection. But they require an extra fine-tuning stage, and the disjointed optimization of representation and classifier might lead to suboptimal results. However, end-to-end training methods, like equalization loss (EQL), still perform worse than decoupled training methods. In this paper, we reveal the main issue in long-tailed object detection is the imbalanced gradients between positives and negatives, and find that EQL does not solve it well. To address the problem of imbalanced gradients, we introduce a new version of equalization loss, called equalization loss v2 (EQL v2), a novel gradient guided reweighing mechanism that re-balances the training process for each category independently and equally. Extensive experiments are performed on the challenging LVIS benchmark. EQL v2 outperforms origin EQL by about 4 points overall AP with 14 - 18 points improvements on the rare categories. More importantly, it also surpasses decoupled training methods. Without further tuning for the Open Images dataset, EQL v2 improves EQL by 7.3 points AP, showing strong generalization ability. Codes have been released at https://github.com/tztztztztz/eqlv2
****
Dynamic Metric Learning: Towards a Scalable Metric Space To Accommodate Multiple Semantic Scales
Yifan Sun, Yuke Zhu, Yuhan Zhang, Pengkun Zheng, Xi Qiu, Chi Zhang, Yichen Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5393-5402
This paper introduces a new fundamental characteristics, i.e., the dynamic range, from real-world metric tools to deep visual recognition. In metrology, the dynamic range is a basic quality of a metric tool, indicating its flexibility to accommodate various scales. Larger dynamic range offers higher flexibility. We argue that such flexibility is also important for deep metric learning, because different visual concepts indeed correspond to different semantic scales. Introducing the dynamic range to deep metric learning, we get a novel computer vision task, i.e., the Dynamic Metric Learning. Dynamic Metric Learning aims to learn a scalable metric space to accommodate visual concepts across multiple semantic scales. Based on three different types of images, i.e., vehicle, animal and online products, we construct three datasets for Dynamic Metric Learning. We benchmark these datasets with popular deep metric learning methods and find Dynamic Metric Learning to be very challenging. The major difficulty lies in a conflict between different scales: the discriminative ability under a small scale usually compromises the discriminative ability under a large one, and vice versa. As a minor contribution, we propose Cross-Scale Learning (CSL) to alleviate such conflict. We show that CSL consistently improves the baseline on all the three datasets.
****
Primitive Representation Learning for Scene Text Recognition
Ruijie Yan, Liangrui Peng, Shanyu Xiao, Gang Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 284-293
Scene text recognition is a challenging task due to diverse variations of text instances in natural scene images. Conventional methods based on CNN-RNN-CTC or encoder-decoder with attention mechanism may not fully investigate stable and efficient feature representations for multi-oriented scene texts. In this paper, we propose a primitive representation learning method that aims to exploit intrinsic representations of scene text images. We model elements in feature maps as the nodes of an undirected graph. A pooling aggregator and a weighted aggregator are proposed to learn primitive representations, which are transformed into high-level visual text representations by graph convolutional networks. A Primitive REpresentation learning Network (PREN) is constructed to use the visual text representations for parallel decoding. Furthermore, by integrating visual text representations into an encoder-decoder model with the 2D attention mechanism, we propose a framework called PREN2D to alleviate the misalignment problem in attentio

n-based methods. Experimental results on both English and Chinese scene text recognition tasks demonstrate that PREN keeps a balance between accuracy and efficiency, while PREN2D achieves state-of-the-art performance.
****
RPSRNet: End-to-End Trainable Rigid Point Set Registration Network Using Barnes-Hut 2D-Tree Representation
Sk Aziz Ali, Kerem Kahraman, Gerd Reis, Didier Stricker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13100-13110
We propose RPSRNet - a novel end-to-end trainable deep neural network for rigid point set registration. For this task, we use a novel 2^D-tree representation for the input point sets and a hierarchical deep feature embedding in the neural network. An iterative transformation refinement module in our network boosts the feature matching accuracy in the intermediate stages. We achieve an inference speed of 12-15ms to register a pair of input point clouds as large as 250K. Extensive evaluation on (i) KITTI LiDAR odometry and (ii) ModelNet-40 datasets shows that our method outperforms prior state-of-the-art methods -- e.g., on the KITTI data set, DCP-v2 by1.3 and 1.5 times, and PointNetLK by 1.8 and 1.9 times better rotational and translational accuracy respectively. Evaluation on ModelNet40 shows that RPSRNet is more robust than other benchmark methods when the samples contain a significant amount of noise and other disturbances. RPSRNet accurately registers point clouds with non-uniform sampling densities, e.g., LiDAR data, which cannot be processed by many existing deep-learning-based registration methods.
****
On the Difficulty of Membership Inference Attacks
Shahbaz Rezaei, Xin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7892-7900
Recent studies propose membership inference (MI) attacks on deep models, where the goal is to infer if a sample has been used in the training process. Despite their apparent success, these studies only report accuracy, precision, and recall of the positive class (member class). Hence, the performance of these attacks have not been clearly reported on negative class (non-member class). In this paper, we show that the way the MI attack performance has been reported is often misleading because they suffer from high false positive rate or false alarm rate (FAR) that has not been reported. FAR shows how often the attack model mislabel non-training samples (non-member) as training (member) ones. The high FAR makes MI attacks fundamentally impractical, which is particularly more significant for tasks such as membership inference where the majority of samples in reality belong to the negative (non-training) class. Moreover, we show that the current MI attack models can only identify the membership of misclassified samples with mediocre accuracy at best, which only constitute a very small portion of training samples. We analyze several new features that have not been comprehensively explored for membership inference before, including distance to the decision boundary and gradient norms, and conclude that deep models' responses are mostly similar among train and non-train samples. We conduct several experiments on image classification tasks, including MNIST, CIFAR-10, CIFAR-100, and ImageNet, using various model architecture, including LeNet, AlexNet, ResNet, etc. We show that the current state-of-the-art MI attacks cannot achieve high accuracy and low FAR at the same time, even when the attacker is given several advantages. The source code is available at https://github.com/shrezaei/MI-Attack.
****
Neural Geometric Level of Detail: Real-Time Rendering With Implicit 3D Shapes
Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11358-11367
Neural signed distance functions (SDFs) are emerging as an effective representation for 3D shapes. State-of-the-art methods typically encode the SDF with a large, fixed-size neural network to approximate complex shapes with implicit surface

s. Rendering with these large networks is, however, computationally expensive since it requires many forward passes through the network for every pixel, making these representations impractical for real-time graphics. We introduce an efficient neural representation that, for the first time, enables real-time rendering of high-fidelity neural SDFs, while achieving state-of-the-art geometry reconstruction quality. We represent implicit surfaces using an octree-based feature volume which adaptively fits shapes with multiple discrete levels of detail (LODs), and enables continuous LOD with SDF interpolation. We further develop an efficient algorithm to directly render our novelneural SDF representation in real-time by querying only the necessary LODswith sparse octree traversal. We show that our representation is 2-3 orders of magnitude more efficient in terms of rendering speed compared to previous works. Furthermore, it produces state-of-the-art reconstruction quality for complex shapes under both 3D geometric and 2D image-space metrics.
****
Pareidolia Face Reenactment
Linsen Song, Wayne Wu, Chaoyou Fu, Chen Qian, Chen Change Loy, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2236-2245
We present a new application direction named Pareidolia Face Reenactment, which is defined as animating a static illusory face to move in tandem with a human face in the video. For the large differences between pareidolia face reenactment and traditional human face reenactment, two main challenges are introduced, i.e., shape variance and texture variance. In this work, we propose a novel Parametric Unsupervised Reenactment Algorithm to tackle these two challenges. Specifically, we propose to decompose the reenactment into three catenate processes: shape modeling, motion transfer and texture synthesis. With the decomposition, we introduce three crucial components, i.e., Parametric Shape Modeling, Expansionary Motion Transfer and Unsupervised Texture Synthesizer, to overcome the problems brought by the remarkably variances on pareidolia faces. Extensive experiments show the superior performance of our method both qualitatively and quantitatively. Code, model and data are available on our project page.
****
ProSelfLC: Progressive Self Label Correction for Training Robust Deep Neural Networks
Xinshao Wang, Yang Hua, Elyor Kodirov, David A. Clifton, Neil M. Robertson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 752-761
To train robust deep neural networks (DNNs), we systematically study several target modification approaches, which include output regularisation, self and non-self label correction (LC). Two key issues are discovered: (1) Self LC is the most appealing as it exploits its own knowledge and requires no extra models. However, how to automatically decide the trust degree of a learner as training goes is not well answered in the literature? (2) Some methods penalise while the others reward low-entropy predictions, prompting us to ask which one is better? To resolve the first issue, taking two well-accepted propositions-deep neural networks learn meaningful patterns before fitting noise (Arpit et al., 2017) and minimum entropy regularisation principle (Grandvalet & Bengio, 2006)-we propose a novel end-to-end method named ProSelfLC, which is designed according to learning time and entropy. Specifically, given a data point, we progressively increase trust in its predicted label distribution versus its annotated one if a model has been trained for enough time and the prediction is of low entropy (high confidence). For the second issue, according to ProSelfLC, we empirically prove that it is better to redefine a meaningful low-entropy status and optimise the learner toward it. This serves as a defence of entropy minimisation. We demonstrate the effectiveness of ProSelfLC through extensive experiments in both clean and noisy settings. The source code is available at https://github.com/XinshaoAmosWang/ProSelfLC-CVPR2021.
****
Learning To Segment Rigid Motions From Two Frames

Gengshan Yang, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1266-1275
Appearance-based detectors achieve remarkable performance on common scenes, benefiting from high-capacity models and massive annotated data, but tend to fail for scenarios that lack training data. Geometric motion segmentation algorithms, however, generalize to novel scenes, but have yet to achieve comparable performance to appearance-based ones, due to noisy motion estimations and degenerate motion configurations. To combine the best of both worlds, we propose a modular network, whose architecture is motivated by a geometric analysis of what independent object motions can be recovered from an ego-motion field. It takes two consecutive frames as input and predicts segmentation masks for the background and multiple rigidly moving objects, which are then parameterized by 3D rigid transformations. Our method achieves state-of-the-art performance for rigid motion segmentation on KITTI and Sintel. The inferred rigid motions lead to a significant improvement for depth and scene flow estimation.
****
Joint Deep Model-Based MR Image and Coil Sensitivity Reconstruction Network (Joint-ICNet) for Fast MRI
Yohan Jun, Hyungseob Shin, Taejoon Eo, Dosik Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5270-5279
Magnetic resonance imaging (MRI) can provide diagnostic information with high-resolution and high-contrast images. However, MRI requires a relatively long scan time compared to other medical imaging techniques, where long scan time might occur patient's discomfort and limit the increase in resolution of magnetic resonance (MR) image. In this study, we propose a Joint Deep Model-based MR Image and Coil Sensitivity Reconstruction Network, called Joint-ICNet, which jointly reconstructs an MR image and coil sensitivity maps from undersampled multi-coil k-space data using deep learning networks combined with MR physical models. Joint-ICNet has two main blocks, where one is an MR image reconstruction block that reconstructs an MR image from undersampled multi-coil k-space data and the other is a coil sensitivity maps reconstruction block that estimates coil sensitivity maps from undersampled multi-coil k-space data. The desired MR image and coil sensitivity maps can be obtained by sequentially estimating them with two blocks based on the unrolled network architecture. To demonstrate the performance of Joint-ICNet, we performed experiments with a fastMRI brain dataset for two reduction factors (R = 4 and 8). With qualitative and quantitative results, we demonstrate that our proposed Joint-ICNet outperforms conventional parallel imaging and deep-learning-based methods in reconstructing MR images from undersampled multi-coil k-space data.
****
On Feature Normalization and Data Augmentation
Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, Kilian Q. Weinberger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12383-12392
The moments (a.k.a., mean and standard deviation) of latent features are often removed as noise when training image recognition models, to increase stability and reduce training time. However, in the field of image generation, the moments play a much more central role. Studies have shown that the moments extracted from instance normalization and positional normalization can roughly capture style and shape information of an image. Instead of being discarded, these moments are instrumental to the generation process. In this paper we propose Moment Exchange, an implicit data augmentation method that encourages the model to utilize the moment information also for recognition models. Specifically, we replace the moments of the learned features of one training image by those of another, and also interpolate the target labels---forcing the model to extract training signal from the moments in addition to the normalized features. As our approach is fast, operates entirely in feature space, and mixes different signals than prior methods, one can effectively combine it with existing augmentation approaches. We demonstrate its efficacy across several recognition benchmark data sets where it im

proves the generalization capability of highly competitive baseline networks wit
h remarkable consistency.
****

SelfDoc: Self-Supervised Document Representation Learning
Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, Hongfu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5652-5660
We propose SelfDoc, a task-agnostic pre-training framework for document image understanding. Because documents are multimodal and are intended for sequential reading, our framework exploits the positional, textual, and visual information of every semantically meaningful component in a document, and it models the contextualization between each block of content. Unlike existing document pre-training models, our model is coarse-grained instead of treating individual words as input, therefore avoiding an overly fine-grained with excessive contextualization. Beyond that, we introduce cross-modal learning in the model pre-training phase to fully leverage multimodal information from unlabeled documents. For downstream usage, we propose a novel modality-adaptive attention mechanism for multimodal feature fusion by adaptively emphasizing language and vision signals. Our framework benefits from self-supervised pre-training on documents without requiring annotations by a feature masking training strategy. It achieves superior performance on multiple downstream tasks with significantly fewer document images used in the pre-training stage compared to previous works.
****

Towards Rolling Shutter Correction and Deblurring in Dynamic Scenes
Zhihang Zhong, Yinqiang Zheng, Imari Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9219-9228
Joint rolling shutter correction and deblurring (RSCD) techniques are critical for the prevalent CMOS cameras. However, current approaches are still based on conventional energy optimization and are developed for static scenes. To enable learning-based approaches to address real-world RSCD problem, we contribute the first dataset, BS-RSCD, which includes both ego-motion and object-motion in dynamic scenes. Real distorted and blurry videos with corresponding ground truth are recorded simultaneously via a beam-splitter-based acquisition system. Since direct application of existing individual rolling shutter correction (RSC) or global shutter deblurring (GSD) methods on RSCD leads to undesirable results due to inherent flaws in the network architecture, we further present the first learning-based model (JCD) for RSCD. The key idea is that we adopt bi-directional warping streams for displacement compensation, while also preserving the non-warped deblurring stream for details restoration. The experimental results demonstrate that JCD achieves state-of-the-art performance on the realistic RSCD dataset (BS-RSCD) and the synthetic RSC dataset (Fastec-RS). The dataset and code are available at https://github.com/zzh-tech/RSCD.
****

VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild
Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4133-4143
In this paper, we present a new dataset with the target of advancing the scene parsing task from images to videos. Our dataset aims to perform Video Scene Parsing in the Wild (VSPW), which covers a wide range of real-world scenarios and categories. To be specific, our VSPW is featured from the following aspects: 1) Well-trimmed long-temporal clips. Each video contains a complete shot, lasting around 5 seconds on average. 2) Dense annotation. The pixel-level annotations are provided at a high frame rate of 15 f/s. 3) High resolution. Over 96% of the captured videos are with high spatial resolutions from 720P to 4K. We totally annotate 3,337 videos, including 239,934 frames from 124 categories. To the best of our knowledge, our VSPW is the first attempt to tackle the challenging video scene parsing task in the wild by considering diverse scenarios. Based on VSPW, we design a generic Temporal Context Blending (TCB) network, which can effectively harness long-range contextual information from the past frames to help segment the

current one. Extensive experiments show that our TCB network improves both the s egmentation performance and temporal stability comparing with image-/video-based state-of-the-art methods. We hope that the scale, diversity, long-temporal, and high frame rate of our VSPW can significantly advance the research of video sce ne parsing and beyond.
****
Multi-Label Learning From Single Positive Labels
Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, Nebojs a Jojic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2021, pp. 933-942
Predicting all applicable labels for a given image is known as multi-label class ification. Compared to the standard multi-class case (where each image has only one label), it is considerably more challenging to annotate training data for mu lti-label classification. When the number of potential labels is large, human an notators find it difficult to mention all applicable labels for each training im age. Furthermore, in some settings detection is intrinsically difficult e.g. fin ding small object instances in high resolution images. As a result, multi-label training data is often plagued by false negatives. We consider the hardest versi on of this problem, where annotators provide only one relevant label for each im age. As a result, training sets will have only one positive label per image and no confirmed negatives. We explore this special case of learning from missing la bels across four different multi-label image classification datasets for both li near classifiers and end-to-end fine-tuned deep networks. We extend existing mul ti-label losses to this setting and propose novel variants that constrain the nu mber of expected positive labels during training. Surprisingly, we show that in some cases it is possible to approach the performance of fully labeled classifie rs despite training with significantly fewer confirmed labels.
****
Towards Part-Based Understanding of RGB-D Scans
Alexey Bokhovkin, Vladislav Ishimtsev, Emil Bogomolov, Denis Zorin, Alexey Artem ov, Evgeny Burnaev, Angela Dai; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2021, pp. 7484-7494
Recent advances in 3D semantic scene understanding have shown impressive progres s in 3D instance segmentation, enabling object-level reasoning about 3D scenes; however, a finer-grained understanding is required to enable interactions with o bjects and their functional understanding. Thus, we propose the task of part-bas ed scene understanding of real-world 3D environments: from an RGB-D scan of a sc ene, we detect objects, and for each object predict its decomposition into geome tric part masks, which composed together form the complete geometry of the obser ved object. We leverage an intermediary part graph representation to enable robu st completion as well as building of part priors, which we use to construct the final part mask predictions. Our experiments demonstrate that guiding part under standing through part graph to part prior-based predictions significantly outper forms alternative approaches to the task of part-based instance completion.
****
Learning Semantic-Aware Dynamics for Video Prediction
Xinzhu Bei, Yanchao Yang, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 902-912
We propose an architecture and training scheme to predict video frames by explic itly modeling dis-occlusions and capturing the evolution of semantically consist ent regions in the video. The scene layout (semantic map) and motion (optical fl ow) are decomposed into layers, which are predicted and fused with their context to generate future layouts and motions. The appearance of the scene is warped f rom past frames using the predicted motion in co-visible regions; dis-occluded r egions are synthesized with content-aware inpainting utilizing the predicted sce ne layout. The result is a predictive model that explicitly represents objects a nd learns their class-specific motion, which we evaluate on video prediction ben chmarks.
****
Bipartite Graph Network With Adaptive Message Passing for Unbiased Scene Graph G

eneration
Rongjie Li, Songyang Zhang, Bo Wan, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11109-11119
Scene graph generation is an important visual understanding task with a broad range of vision applications. Despite recent tremendous progress, it remains challenging due to the intrinsic long-tailed class distribution and large intra-class variation. To address these issues, we introduce a novel confidence-aware bipartite graph neural network with adaptive message propagation mechanism for unbiased scene graph generation. In addition, we propose an efficient bi-level data resampling strategy to alleviate the imbalanced data distribution problem in training our graph network. Our approach achieves superior or competitive performance over previous methods on several challenging datasets, including Visual Genome, Open Images V4/V6, demonstrating its effectiveness and generality.
****
Guided Interactive Video Object Segmentation Using Reliability-Based Attention Maps
Yuk Heo, Yeong Jun Koh, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7322-7330
We propose a novel guided interactive segmentation (GIS) algorithm for video objects to improve the segmentation accuracy and reduce the interaction time. First, we design the reliability-based attention module to analyze the reliability of multiple annotated frames. Second, we develop the intersection-aware propagation module to propagate segmentation results to neighboring frames. Third, we introduce the GIS mechanism for a user to select unsatisfactory frames quickly with less effort. Experimental results demonstrate that the proposed algorithm provides more accurate segmentation results at a faster speed than conventional algorithms. Codes are available at https://github.com/yuk6heo/GIS-RAmap.
****
Learning Spatial-Semantic Relationship for Facial Attribute Recognition With Limited Labeled Data
Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, Hanzi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11916-11925
Recent advances in deep learning have demonstrated excellent results for Facial Attribute Recognition (FAR), typically trained with large-scale labeled data. However, in many real-world FAR applications, only limited labeled data are available, leading to remarkable deterioration in performance for most existing deep learning-based FAR methods. To address this problem, here we propose a method termed Spatial-Semantic Patch Learning (SSPL). The training of SSPL involves two stages. First, three auxiliary tasks, consisting of a Patch Rotation Task (PRT), a Patch Segmentation Task (PST), and a Patch Classification Task (PCT), are jointly developed to learn the spatial-semantic relationship from large-scale unlabeled facial data. We thus obtain a powerful pre-trained model. In particular, PRT exploits the spatial information of facial images in a self-supervised learning manner. PST and PCT respectively capture the pixel-level and image-level semantic information of facial images based on a facial parsing model. Second, the spatial-semantic knowledge learned from auxiliary tasks is transferred to the FAR task. By doing so, it enables that only a limited number of labeled data are required to fine-tune the pre-trained model. We achieve superior performance compared with state-of-the-art methods, as substantiated by extensive experiments and studies.
****
Decoupled Dynamic Filter Networks
Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6647-6656
Convolution is one of the basic building blocks of CNN architectures. Despite its common use, standard convolution has two main shortcomings: Content-agnostic and Computation-heavy. Dynamic filters are content-adaptive, while further increasing the computational overhead. Depth-wise convolution is a lightweight variant

, but it usually leads to a drop in CNN performance or requires a larger number of channels. In this work, we propose the Decoupled Dynamic Filter (DDF) that ca n simultaneously tackle both of these shortcomings. Inspired by recent advances in attention, DDF decouples a depth-wise dynamic filter into spatial and channel dynamic filters. This decomposition considerably reduces the number of paramete rs and limits computational costs to the same level as depth-wise convolution. M eanwhile, we observe a significant boost in performance when replacing standard convolution with DDF in classification networks. ResNet50 / 101 get improved by 1.9% and 1.3% on the top-1 accuracy, while their computational costs are reduced by nearly half. Experiments on the detection and joint upsampling networks also demonstrate the superior performance of the DDF upsampling variant (DDF-Up) in comparison with standard convolution and specialized content-adaptive layers.
****
Motion Representations for Articulated Animation
Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, Sergey Tulyakov ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2021, pp. 13653-13662
We propose novel motion representations for animating articulated objects consis ting of distinct parts. In a completely unsupervised manner, our method identifi es object parts, tracks them in a driving video, and infers their motions by con sidering their principal axes. In contrast to the previous keypoint-based works, our method extracts meaningful and consistent regions, describing locations, sh ape, and pose. The regions correspond to semantically relevant and distinct obje ct parts, that are more easily detected in frames of the driving video. To force decoupling of foreground from background, we model non-object related global mo tion with an additional affine transformation. To facilitate animation and preve nt the leakage of the shape of the driving object, we disentangle shape and pose of objects in the region space. Our model can animate a variety of objects, sur passing previous methods by a large margin on existing benchmarks. We present a challenging new benchmark with high-resolution videos and show that the improvem ent is particularly pronounced when articulated objects are considered, reaching 96.6% user preference vs. the state of the art.
****
General Multi-Label Image Classification With Transformers
Jack Lanchantin, Tianlu Wang, Vicente Ordonez, Yanjun Qi; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16 478-16488
Multi-label image classification is the task of predicting a set of labels corre sponding to objects, attributes or other entities present in an image. In this w ork we propose the Classification Transformer (C-Tran), a general framework for multi-label image classification that leverages Transformers to exploit the comp lex dependencies among visual features and labels. Our approach consists of a Tr ansformer encoder trained to predict a set of target labels given an input set o f masked labels, and visual features from a convolutional neural network. A key ingredient of our method is a label mask training objective that uses a ternary encoding scheme to represent the state of the labels as positive, negative, or u nknown during training. Our model shows state-of-the-art performance on challeng ing datasets such as COCO and Visual Genome. Moreover, because our model explici tly represents the label state during training, it is more general by allowing u s to produce improved results for images with partial or extra label annotations during inference. We demonstrate this additional capability in the COCO, Visual Genome, News-500, and CUB image datasets.
****
On Self-Contact and Human Pose
Lea Muller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, Michael J. Black; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9990-9999
People touch their face 23 times an hour, they cross their arms and legs, put th eir hands on their hips, etc. While many images of people contain some form of s elf-contact, current 3D human pose and shape (HPS) regression methods typically

fail to estimate this contact. To address this, we develop new datasets and meth ods that significantly improve human pose estimation with self-contact. First, w e create a dataset of 3D Contact Poses (3DCP) containing SMPL-X bodies fit to 3D scans as well as poses from AMASS, which we refine to ensure good contact. Seco nd, we leverage this to create the Mimic-The-Pose (MTP) dataset of images, colle cted via Amazon Mechanical Turk, containing people mimicking the 3DCP poses with self-contact. Third, we develop a novel HPS optimization method, SMPLify-XMC, t hat includes contact constraints and uses the known 3DCP body pose during fittin g to create near ground-truth poses for MTP images. Fourth, for more image varie ty, we label a dataset of in-the-wild images with Discrete Self-Contact (DSC) in formation and use another new optimization method, SMPLify-DC, that exploits dis crete contacts during pose optimization. Finally, we use our datasets during SPI N training to learn a new 3D human pose regressor, called TUCH (Towards Understa nding Contact in Humans). We show that the new self-contact training data signif icantly improves 3D human pose estimates on withheld test data and existing data sets like 3DPW. Not only does our method improve results for self-contact poses, but it also improves accuracy for non-contact poses. The code and data are avai lable for research purposes at https://tuch.is.tue.mpg.de.
****
Center-Based 3D Object Detection and Tracking
Tianwei Yin, Xingyi Zhou, Philipp Krahenbuhl; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11784-11793
Three-dimensional objects are commonly represented as 3D boxes in a point-cloud. This representation mimics the well-studied image-based 2D bounding-box detecti on but comes with additional challenges. Objects in a 3D world do not follow any particular orientation, and box-based detectors have difficulties enumerating a ll orientations or fitting an axis-aligned bounding box to rotated objects. In t his paper, we instead propose to represent, detect, and track 3D objects as poin ts. Our framework, CenterPoint, first detects centers of objects using a keypoin t detector and regresses to other attributes, including 3D size, 3D orientation, and velocity. In a second stage, it refines these estimates using additional po int features on the object. In CenterPoint, 3D object tracking simplifies to gre edy closest-point matching. The resulting detection and tracking algorithm is si mple, efficient, and effective. On the nuScenes and Waymo datasets, CenterPoint surpasses prior methods by a large margin. On the Waymo Open Dataset, CenterPoin t improves previous state-of-the-art by 10-20% while running at 13FPS. The code and pretrained models are available at https://github.com/tianweiy/CenterPoint.
****
Prototype Augmentation and Self-Supervision for Incremental Learning
Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, Cheng-Lin Liu; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5871-5880
Despite the impressive performance in many individual tasks, deep neural network s suffer from catastrophic forgetting when learning new tasks incrementally. Rec ently, various incremental learning methods have been proposed, and some approac hes achieved acceptable performance relying on stored data or complex generative models. However, storing data from previous tasks is limited by memory or priva cy issues, and generative models are usually unstable and inefficient in trainin g. In this paper, we propose a simple non-exemplar based method named PASS, to a ddress the catastrophic forgetting problem in incremental learning. On the one h and, we propose to memorize one class-representative prototype for each old clas s and adopt prototype augmentation (protoAug) in the deep feature space to maint ain the decision boundary of previous tasks. On the other hand, we employ self-s upervised learning (SSL) to learn more generalizable and transferable features f or other tasks, which demonstrates the effectiveness of SSL in incremental learn ing. Experimental results on benchmark datasets show that our approach significa ntly outperforms non-exemplar based methods, and achieves comparable performance compared to exemplar based approaches.
****
CompositeTasking: Understanding Images by Spatial Composition of Tasks

Nikola Popovic, Danda Pani Paudel, Thomas Probst, Guolei Sun, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6870-6880
We define the concept of CompositeTasking as the fusion of multiple, spatially distributed tasks, for various aspects of image understanding. Learning to perform spatially distributed tasks is motivated by the frequent availability of only sparse labels across tasks, and the desire for a compact multi-tasking network. To facilitate CompositeTasking, we introduce a novel task conditioning model -- a single encoder-decoder network that performs multiple, spatially varying tasks at once. The proposed network takes an image and a set of pixel-wise dense task requests as inputs, and performs the requested prediction task for each pixel. Moreover, we also learn the composition of tasks that needs to be performed according to some CompositeTasking rules, which includes the decision of where to apply which task. It not only offers us a compact network for multi-tasking, but also allows for task-editing. Another strength of the proposed method is demonstrated by only having to supply sparse supervision per task. The obtained results are on par with our baselines that use dense supervision and a multi-headed multi-tasking design. The source code will be made publicly available at www.github.com/nikola3794/composite-tasking.
****
Searching for Fast Model Families on Datacenter Accelerators
Sheng Li, Mingxing Tan, Ruoming Pang, Andrew Li, Liqun Cheng, Quoc V. Le, Norman P. Jouppi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8085-8095
Neural Architecture Search (NAS), together with model scaling, has shown remarkable progress in designing high accuracy and fast convolutional architecture families. However, as neither NAS nor model scaling considers sufficient hardware architecture details, they do not take full advantage of the emerging datacenter (DC) accelerators. In this paper, we search for fast and accurate CNN model families for efficient inference on DC accelerators. We first analyze DC accelerators and find that existing CNNs suffer from insufficient operational intensity, parallelism, and execution efficiency and exhibit FLOPs-latency nonproportionality. These insights let us create a DC-accelerator-optimized search space, with space-to-depth, space-to-batch, hybrid fused convolution structures with vanilla and depthwise convolutions, and block-wise activation functions. We further propose a latency-aware compound scaling (LACS), the first multi-objective compound scaling method optimizing both accuracy and latency. Our LACS discovers that network depth should grow much faster than image size and network width, which is quite different from the observations from previous compound scaling. With the new search space and LACS, our search and scaling on datacenter accelerators results in a new model series named EfficientNet-X. EfficientNet-X is up to more than 2X faster than EfficientNet (a model series with state-of-the-art trade-off on FLOPs and accuracy) on TPUv3 and GPUv100, with comparable accuracy. EfficientNet-X is also up to 7X faster than recent RegNet and ResNeSt on TPUv3 and GPUv100. Source code is at https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet/tpu
****
Task-Aware Variational Adversarial Active Learning
Kwanyoung Kim, Dongwon Park, Kwang In Kim, Se Young Chun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8166-8175
Often, labeling large amount of data is challenging due to high labeling cost limiting the application domain of deep learning techniques. Active learning (AL) tackles this by querying the most informative samples to be annotated among unlabeled pool. Two promising directions for AL that have been recently explored are task-agnostic approach to select data points that are far from the current labeled pool and task-aware approach that relies on the perspective of task model. Unfortunately, the former does not exploit structures from tasks and the latter does not seem to well-utilize overall data distribution. Here, we propose task-aware variational adversarial AL (TA-VAAL) that modifies task-agnostic VAAL, that

considered data distribution of both label and unlabeled pools, by relaxing task learning loss prediction to ranking loss prediction and by using ranking conditional generative adversarial network to embed normalized ranking loss information on VAAL. Our proposed TA-VAAL outperforms state-of-the-arts on various benchmark datasets for classifications with balanced / imbalanced labels as well as semantic segmentation and its task-aware and task-agnostic AL properties were confirmed with our in-depth analyses.
****
Understanding and Simplifying Perceptual Distances
Dan Amir, Yair Weiss; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12226-12235
Perceptual metrics based on features of deep Convolutional Neural Networks (CNNs) have shown remarkable success when used as loss functions in a range of computer vision problems and significantly outperform classical losses such as L1 or L2 in pixel space. The source of this success remains somewhat mysterious, especially since a good loss does not require a particular CNN architecture nor a particular training method. In this paper we show that similar success can be achieved even with losses based on features of a deep CNN with random filters. We use the tool of infinite CNNs to derive an analytical form for perceptual similarity in such CNNs, and prove that the perceptual distance between two images is equivalent to the maximum mean discrepancy (MMD) distance between local distributions of small patches in the two images. We use this equivalence to propose a simple metric for comparing two images which directly computes the MMD between local distributions of patches in the two images. Our proposed metric is simple to understand, requires no deep networks, and gives comparable performance to perceptual metrics in a range of computer vision tasks.
****
Class-Aware Robust Adversarial Training for Object Detection
Pin-Chun Chen, Bo-Han Kung, Jun-Cheng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10420-10429
Object detection is an important computer vision task with plenty of real-world applications; therefore, how to enhance its robustness against adversarial attacks has emerged as a crucial issue. However, most of the previous defense methods focused on the classification task and had few analysis in the context of the object detection task. In this work, to address the issue, we present a novel class-aware robust adversarial training paradigm for the object detection task. For a given image, the proposed approach generates a universal adversarial perturbation to simultaneously attack all the occurred objects in the image through jointly maximizing the respective loss for each object. Meanwhile, instead of normalizing the total loss with the number of objects, the proposed approach decomposes the total loss into class-wise losses and normalizes each class loss using the number of objects for the class. The adversarial training based on the class weighted loss can not only balances the influence of each class but also effectively and evenly improves the adversarial robustness of trained models for all the object classes as compared with the previous defense methods. Furthermore, with the recent development of fast adversarial training, we provide a fast version of the proposed algorithm which can be trained faster than the traditional adversarial training while keeping comparable performance. With extensive experiments on the challenging PASCAL-VOC and MS-COCO datasets, the evaluation results demonstrate that the proposed defense methods can effectively enhance the robustness of the object detection models.
****
Bayesian Nested Neural Networks for Uncertainty Calibration and Adaptive Compression
Yufei Cui, Ziquan Liu, Qiao Li, Antoni B. Chan, Chun Jason Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2392-2401
Nested networks or slimmable networks are neural networks whose architectures can be adjusted instantly during testing time, e.g., based on computational constraints. Recent studies have focused on a "nested dropout" layer, which is able to

order the nodes of a layer by importance during training, thus generating a nested set of sub-networks that are optimal for different configurations of resources. However, the dropout rate is fixed as a hyper-parameter over different layers during the whole training process. Therefore, when nodes are removed, the performance decays in a human-specified trajectory rather than in a trajectory learned from data. Another drawback is the generated sub-networks are deterministic networks without well-calibrated uncertainty. To address these two problems, we develop a Bayesian approach to nested neural networks. We propose a variational ordering unit that draws samples for nested dropout at a low cost, from a proposed Downhill distribution, which provides useful gradients to the parameters of nested dropout. Based on this approach, we design a Bayesian nested neural network that learns the order knowledge of the node distributions. In experiments, we show that the proposed approach outperforms the nested network in terms of accuracy, calibration, and out-of-domain detection in classification tasks. It also outperforms the related approach on uncertainty-critical tasks in computer vision.
****

Fast Bayesian Uncertainty Estimation and Reduction of Batch Normalized Single Image Super-Resolution Network
Aupendu Kar, Prabir Kumar Biswas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4957-4966
Convolutional neural network (CNN) has achieved unprecedented success in image super-resolution tasks in recent years. However, the network's performance depends on the distribution of the training sets and degrades on out-of-distribution samples. This paper adopts a Bayesian approach for estimating uncertainty associated with output and applies it in a deep image super-resolution model to address the concern mentioned above. We use the uncertainty estimation technique using the batch-normalization layer, where stochasticity of the batch mean and variance generate Monte-Carlo (MC) samples. The MC samples, which are nothing but different super-resolved images using different stochastic parameters, reconstruct the image, and provide a confidence or uncertainty map of the reconstruction. We propose a faster approach for MC sample generation, and it allows the variable image size during testing. Therefore, it will be useful for image reconstruction domain. Our experimental findings show that this uncertainty map strongly relates to the quality of reconstruction generated by the deep CNN model and explains its limitation. Furthermore, this paper proposes an approach to reduce the model's uncertainty for an input image, and it helps to defend the adversarial attacks on the image super-resolution model. The proposed uncertainty reduction technique also improves the performance of the model for out-of-distribution test images. To the best of our knowledge, we are the first to propose an adversarial defense mechanism in any image reconstruction domain.
****

Euro-PVI: Pedestrian Vehicle Interactions in Dense Urban Centers
Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6408-6417
Accurate prediction of pedestrian and bicyclist paths is integral to the development of reliable autonomous vehicles in dense urban environments. The interactions between vehicle and pedestrian or bicyclist have a significant impact on the trajectories of traffic participants e.g. stopping or turning to avoid collisions. Although recent datasets and trajectory prediction approaches have fostered the development of autonomous vehicles yet the amount of vehicle-pedestrian (bicyclist) interactions modeled are sparse. In this work, we propose Euro-PVI, a dataset of pedestrian and bicyclist trajectories. In particular, our dataset caters more diverse and complex interactions in dense urban scenarios compared to the existing datasets. To address the challenges in predicting future trajectories with dense interactions, we develop a joint inference model that learns an expressive multi-modal shared latent space across agents in the urban scene. This enables our Joint-b-cVAE approach to better model the distribution of future trajectories. We achieve state of the art results on the nuScenes and Euro-PVI datasets demonstrating the importance of capturing interactions between ego-vehicle and

pedestrians (bicyclists) for accurate predictions.
****
RepVGG: Making VGG-Style ConvNets Great Again
Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13733-13742
We present a simple but powerful architecture of convolutional neural network, which has a VGG-like inference-time body composed of nothing but a stack of 3x3 convolution and ReLU, while the training-time model has a multi-branch topology. Such decoupling of the training-time and inference-time architecture is realized by a structural re-parameterization technique so that the model is named RepVGG. On ImageNet, RepVGG reaches over 80% top-1 accuracy, which is the first time for a plain model, to the best of our knowledge. On NVIDIA 1080Ti GPU, RepVGG models run 83% faster than ResNet-50 or 101% faster than ResNet-101 with higher accuracy and show favorable accuracy-speed trade-off compared to the state-of-the-art models like EfficientNet and RegNet. The code and trained models are available at https://github.com/megvii-model/RepVGG.
****
Partial Feature Selection and Alignment for Multi-Source Domain Adaptation
Yangye Fu, Ming Zhang, Xing Xu, Zuo Cao, Chao Ma, Yanli Ji, Kai Zuo, Huimin Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16654-16663
Multi-Source Domain Adaptation (MSDA), which dedicates to transfer the knowledge learned from multiple source domains to an unlabeled target domain, has drawn increasing attention in the research community. By assuming that the source and target domains share consistent key feature representations and identical label space, existing studies on MSDA typically utilize the entire union set of features from both the source and target domains to obtain the feature map and align the map for each category and domain. However, the default setting of MSDA may neglect the issue of "partialness", i.e., 1) a part of the features contained in the union set of multiple source domains may not present in the target domain; 2) the label space of the target domain may not completely overlap with the multiple source domains. In this paper, we unify the above two cases to a more generalized MSDA task as Multi-Source Partial Domain Adaptation (MSPDA). We propose a novel model termed Partial Feature Selection and Alignment (PFSA) to jointly cope with both MSDA and MSPDA tasks. Specifically, we firstly employ a feature selection vector based on the correlation among the features of multiple sources and target domains. We then design three effective feature alignment losses to jointly align the selected features by preserving the domain information of the data sample clusters in the same category and the discrimination between different classes. Extensive experiments on various benchmark datasets for both MSDA and MSPDA tasks demonstrate that our proposed PFSA approach remarkably outperforms the state-of-the-art MSDA and unimodal PDA methods.
****
Multi-Institutional Collaborations for Improving Deep Learning-Based Magnetic Resonance Image Reconstruction Using Federated Learning
Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2423-2432
Fast and accurate reconstruction of magnetic resonance (MR) images from under-sampled data is important in many clinical applications. In recent years, deep learning-based methods have been shown to produce superior performance on MR image reconstruction. However, these methods require large amounts of data which is difficult to collect and share due to the high cost of acquisition and medical data privacy regulations. In order to overcome this challenge, we propose a federated learning (FL) based solution in which we take advantage of the MR data available at different institutions while preserving patients' privacy. However, the generalizability of models trained with the FL setting can still be suboptimal due to domain shift, which results from the data collected at multiple institutions with different sensors, disease types, and acquisition protocols, etc. With th

e motivation of circumventing this challenge, we propose a cross-site modeling for MR image reconstruction in which the learned intermediate latent features among different source sites are aligned with the distribution of the latent features at the target site. Extensive experiments are conducted to provide various insights about FL for MR image reconstruction. Experimental results demonstrate that the proposed framework is a promising direction to utilize multi-institutional data without compromising patients' privacy for achieving improved MR image reconstruction. Our code is available at https://github.com/guopengf/FL-MRCM
****
UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles
Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, Zhiheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16266-16275
Human behavior understanding with unmanned aerial vehicles (UAVs) is of great significance for a wide range of applications, which simultaneously brings an urgent demand of large, challenging, and comprehensive benchmarks for the development and evaluation of UAV-based models. However, existing benchmarks have limitations in terms of the amount of captured data, types of data modalities, categories of provided tasks, and diversities of subjects and environments. Here we propose a new benchmark - UAV-Human - for human behavior understanding with UAVs, which contains 67,428 multi-modal video sequences and 119 subjects for action recognition, 22,476 frames for pose estimation, 41,290 frames and 1,144 identities for person re-identification, and 22,263 frames for attribute recognition. Our dataset was collected by a flying UAV in multiple urban and rural districts in both daytime and nighttime over three months, hence covering extensive diversities w.r.t subjects, backgrounds, illuminations, weathers, occlusions, camera motions, and UAV flying attitudes. Such a comprehensive and challenging benchmark shall be able to promote the research of UAV-based human behavior understanding, including action recognition, pose estimation, re-identification, and attribute recognition. Furthermore, we propose a fisheye-based action recognition method that mitigates the distortions in fisheye videos via learning unbounded transformations guided by flat RGB videos. Experiments show the efficacy of our method on the UAV-Human dataset.
****
An Alternative Probabilistic Interpretation of the Huber Loss
Gregory P. Meyer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5261-5269
The Huber loss is a robust loss function used for a wide range of regression tasks. To utilize the Huber loss, a parameter that controls the transitions from a quadratic function to an absolute value function needs to be selected. We believe the standard probabilistic interpretation that relates the Huber loss to the Huber density fails to provide adequate intuition for identifying the transition point. As a result, a hyper-parameter search is often necessary to determine an appropriate value. In this work, we propose an alternative probabilistic interpretation of the Huber loss, which relates minimizing the loss to minimizing an upper-bound on the Kullback-Leibler divergence between Laplace distributions, where one distribution represents the noise in the ground-truth and the other represents the noise in the prediction. In addition, we show that the parameters of the Laplace distributions are directly related to the transition point of the Huber loss. We demonstrate, through a toy problem, that the optimal transition point of the Huber loss is closely related to the distribution of the noise in the ground-truth data. As a result, our interpretation provides an intuitive way to identify well-suited hyper-parameters by approximating the amount of noise in the data, which we demonstrate through a case study and experimentation on the Faster R-CNN and RetinaNet object detectors.
****
Siamese Natural Language Tracker: Tracking by Natural Language Descriptions With Siamese Trackers
Qi Feng, Vitaly Ablavsky, Qinxun Bai, Stan Sclaroff; Proceedings of the IEEE/CVF

 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5851-58
60
We propose a novel Siamese Natural Language Tracker (SNLT), which brings the adv
ancements in visual tracking to the tracking by natural language (NL) specificat
ion task. The proposed SNLT is applicable to a wide range of Siamese trackers, p
roviding a new class of baselines for the tracking by NL task and promising futu
re improvements from the advancements of Siamese trackers. The carefully designe
d architecture of the Siamese Natural Language Region Proposal Network (SNL-RPN)
, together with the Dynamic Aggregation of vision and language modalities, is in
troduced to perform the tracking by NL task. Empirical results over tracking ben
chmarks with NL annotations show that the proposed SNLT improves Siamese tracker
s by 3 to 7 percentage points with a slight tradeoff of speed. The proposed SNLT
 outperforms all NL trackers to-date and is competitive among state-of-the-art r
eal-time trackers on LaSOT benchmarks while running at 50 frames per second on a
 single GPU.
****
Discrimination-Aware Mechanism for Fine-Grained Representation Learning
Furong Xu, Meng Wang, Wei Zhang, Yuan Cheng, Wei Chu; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 813-82
2
Recently, with the emergence of retrieval requirements for certain individual in
 the same superclass, e.g., birds, persons, cars, fine-grained recognition task
has attracted a significant amount of attention from academia and industry. In f
ine-grained recognition scenario, the inter-class differences are quite diverse
and subtle, which makes it challenging to extract all the discriminative cues. T
raditional training mechanism optimizes the overall discriminativeness of the wh
ole feature. It may stop early when some feature elements has been trained to di
stinguish training samples well, leaving other elements insufficiently trained f
or a feature. This would result in a less generalizable feature extractor that o
nly captures major discriminative cues and ignores subtle ones. Therefore, there
 is a need for a training mechanism that enforces the discriminativeness of all
the elements in the feature to capture more the subtle visual cues. In this pape
r, we propose a Discrimination-Aware Mechanism (DAM) that iteratively identifies
 insufficiently trained elements and improves them. DAM is able to increase the
number of well learned elements, which captures more visual cues by the feature
extractor. In this way, a more informative representation is learned, which brin
gs better generalization performance. We show that DAM can be easily applied to
both proxy-based and pair-based loss functions, and thus can be used in most exi
sting fine-grained recognition paradigms. Comprehensive experiments on CUB-200-2
011, Cars196, Market-1501, and MSMT17 datasets demonstrate the advantages of our
 DAM based loss over the related state-of-the-art approaches.
****
Rainbow Memory: Continual Learning With a Memory of Diverse Samples
Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, Jonghyun Choi; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 8218-8227
Continual learning is a realistic learning scenario for AI models. Prevalent sce
nario of continual learning, however, assumes disjoint sets of classes as tasks
and is less realistic rather artificial. Instead, we focus on 'blurry' task boun
dary; where tasks shares classes and is more realistic and practical. To address
 such task, we argue the importance of diversity of samples in an episodic memor
y. To enhance the sample diversity in the memory, we propose a novel memory mana
gement strategy based on per-sample classification uncertainty and data augmenta
tion, named Rainbow Memory (RM). With extensive empirical validations on MNIST,
CIFAR10, CIFAR100, and ImageNet datasets, we show that the proposed method signi
ficantly improves the accuracy in blurry continual learning setups, outperformin
g state of the arts by large margins despite its simplicity. Code and data split
s will be available in https://github.com/clovaai/rainbow-memory.
****
Learning Discriminative Prototypes With Dynamic Time Warping

Xiaobin Chang, Frederick Tung, Greg Mori; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8395-8404
Dynamic Time Warping (DTW) is widely used for temporal data processing. However, existing methods can neither learn the discriminative prototypes of different classes nor exploit such prototypes for further analysis. We propose Discriminative Prototype DTW (DP-DTW), a novel method to learn class-specific discriminative prototypes for temporal recognition tasks. DP-DTW shows superior performance compared to conventional DTWs on time series classification benchmarks. Combined with end-to-end deep learning, DP-DTW can handle challenging weakly supervised action segmentation problems and achieves state of the art results on standard benchmarks. Moreover, detailed reasoning on the input video is enabled by the learned action prototypes. Specifically, an action-based video summarization can be obtained by aligning the input sequence with action prototypes.
****
Deep Implicit Moving Least-Squares Functions for 3D Reconstruction
Shi-Lin Liu, Hao-Xiang Guo, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1788-1797
Point set is a flexible and lightweight representation widely used for 3D deep learning. However, their discrete nature prevents them from representing continuous and fine geometry, posing a major issue for learning-based shape generation. In this work, we turn the discrete point sets into smooth surfaces by introducing the well-known implicit moving least-squares (IMLS) surface formulation, which naturally defines locally implicit functions on point sets. We incorporate IMLS surface generation into deep neural networks for inheriting both the flexibility of point sets and the high quality of implicit surfaces. Our IMLSNet predicts an octree structure as a scaffold for generating MLS points where needed and characterizes shape geometry with learned local priors. Furthermore, our implicit function evaluation is independent of the neural network once the MLS points are predicted, thus enabling fast runtime evaluation. Our experiments on 3D object reconstruction demonstrate that IMLSNets outperform state-of-the-art learning-based methods in terms of reconstruction quality and computational efficiency. Extensive ablation tests also validate our network design and loss functions.
****
Video Prediction Recalling Long-Term Motion Context via Memory Alignment Learning
Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, Yong Man Ro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3054-3063
Our work addresses long-term motion context issues for predicting future frames. To predict the future precisely, it is required to capture which long-term motion context (e.g., walking or running) the input motion (e.g., leg movement) belongs to. The bottlenecks arising when dealing with the long-term motion context are: (i) how to predict the long-term motion context naturally matching input sequences with limited dynamics, (ii) how to predict the long-term motion context with high-dimensionality (e.g., complex motion). To address the issues, we propose novel motion context-aware video prediction. To solve the bottleneck (i), we introduce a long-term motion context memory (LMC-Memory) with memory alignment learning. The proposed memory alignment learning enables to store long-term motion contexts into the memory and to match them with sequences including limited dynamics. As a result, the long-term context can be recalled from the limited input sequence. In addition, to resolve the bottleneck (ii), we propose memory query decomposition to store local motion context (i.e., low-dimensional dynamics) and recall the suitable local context for each local part of the input individually. It enables to boost the alignment effects of the memory. Experimental results show that the proposed method outperforms other sophisticated RNN-based methods, especially in long-term condition. Further, we validate the effectiveness of the proposed network designs by conducting ablation studies and memory feature analysis. The source code of this work is available.
****

Automatic Vertebra Localization and Identification in CT by Spine Rectification and Anatomically-Constrained Optimization
Fakai Wang, Kang Zheng, Le Lu, Jing Xiao, Min Wu, Shun Miao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5280-5288
Accurate vertebra localization and identification are required in many clinical applications of spine disorder diagnosis and surgery planning. However, significant challenges are posed in this task by highly varying pathologies (such as vertebral compression fracture, scoliosis, and vertebral fixation) and imaging conditions (such as limited field of view and metal streak artifacts). This paper proposes a robust and accurate method that effectively exploits the anatomical knowledge of the spine to facilitate vertebra localization and identification. A key point localization model is trained to produce activation maps of vertebra centers. They are then re-sampled along the spine centerline to produce spine-rectified activation maps, which are further aggregated into 1-D activation signals. Following this, an anatomically-constrained optimization module is introduced to jointly search for the optimal vertebra centers under a soft constraint that regulates the distance between vertebrae and a hard constraint on the consecutive vertebra indices. When being evaluated on a major public benchmark of 302 highly pathological CT images, the proposed method reports the state of the art identification (id.) rate of 97.4%, and outperforms the best competing method of 94.7% id. rate by reducing the relative id. error rate by half.
****

MotionRNN: A Flexible Model for Video Prediction With Spacetime-Varying Motions
Haixu Wu, Zhiyu Yao, Jianmin Wang, Mingsheng Long; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15435-15444
This paper tackles video prediction from a new dimension of predicting spacetime-varying motions that are incessantly changing across both space and time. Prior methods mainly capture the temporal state transitions but overlook the complex spatiotemporal variations of the motion itself, making them difficult to adapt to ever-changing motions. We observe that physical world motions can be decomposed into transient variation and motion trend, while the latter can be regarded as the accumulation of previous motions. Thus, simultaneously capturing the transient variation and the motion trend is the key to make spacetime-varying motions more predictable. Based on these observations, we propose the MotionRNN framework, which can capture the complex variations within motions and adapt to spacetime-varying scenarios. MotionRNN has two main contributions. The first is that we design the MotionGRU unit, which can model the transient variation and motion trend in a unified way. The second is that we apply the MotionGRU to RNN-based predictive models and indicate a new flexible video prediction architecture with a Motion Highway that can significantly improve the ability to predict changeable motions and avoid motion vanishing for stacked multiple-layer predictive models. With high flexibility, this framework can adapt to a series of models for deterministic spatiotemporal prediction. Our MotionRNN can yield significant improvements on three challenging benchmarks for video prediction with spacetime-varying motions.
****

MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space
Rui Huang, Yixuan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8710-8719
Detecting out-of-distribution (OOD) inputs is a central challenge for safely deploying machine learning models in the real world. Existing solutions are mainly driven by small datasets, with low resolution and very few class labels (e.g., CIFAR). As a result, OOD detection for large-scale image classification tasks remains largely unexplored. In this paper, we bridge this critical gap by proposing a group-based OOD detection framework, along with a novel OOD scoring function termed MOS. Our key idea is to decompose the large semantic space into smaller groups with similar concepts, which allows simplifying the decision boundaries between in- vs. out-of-distribution data for effective OOD detection. Our method s

cales substantially better for high-dimensional class space than previous approa
ches. We evaluate models trained on ImageNet against four carefully curated OOD
datasets, spanning diverse semantics. MOS establishes state-of-the-art performan
ce, reducing the average FPR95 by 14.33% while achieving 6x speedup in inference
 compared to the previous best method.
****
Visual Semantic Role Labeling for Video Understanding
Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, Aniruddha Kembhavi; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2021, pp. 5589-5600
We propose a new framework for understanding and representing related salient ev
ents in a video using visual semantic role labeling. We represent videos as a se
t of related events, wherein each event consists of a verb and multiple entities
 that fulfill various roles relevant to that event. To study the challenging tas
k of semantic role labeling in videos or VidSRL, we introduce the VidSitu benchm
ark, a large scale video understanding data source with 27K 10-second movie clip
s richly annotated with a verb and semantic-roles every 2 seconds. Entities are
co-referenced across events within a movie clip and events are connected to each
 other via event-event relations. Clips in VidSitu are drawn from a large collec
tion of movies (3K) and have been chosen to be both complex (4.2 unique verbs
within a video) as well as diverse (200 verbs have more than 100 annotations ea
ch). We provide a comprehensive analysis of the dataset in comparison to other p
ublicly available video understanding benchmarks, several illustrative baselines
 and evaluate a range of standard video recognition models. Our code and dataset
 will be released publicly.
****
SwiftNet: Real-Time Video Object Segmentation
Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, Song Bai; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
 1296-1305
In this work we present SwiftNet for real-time semi-supervised video object segm
entation (one-shot VOS), which reports 77.8% J&F and 70 FPS on DAVIS 2017 valida
tion dataset, leading all present solutions in overall accuracy and speed perfor
mance. We achieve this by elaborately compressing spatiotemporal redundancy in m
atching-based VOS via Pixel-Adaptive Memory (PAM). Temporally, PAM adaptively tr
iggers memory updates on frames where objects display noteworthy inter-frame var
iations. Spatially, PAM selectively performs memory update and match on dynamic
pixels while ignoring the static ones, significantly reducing redundant computat
ions wasted on segmentation-irrelevant pixels. To promote efficient reference en
coding, light-aggregation encoder is also introduced in SwiftNet deploying rever
sed sub-pixel. We hope SwiftNet could set a strong and efficient baseline for re
al-time VOS and facilitate its application in mobile vision. The source code of
SwiftNet can be found at https://github.com/haochenheheda/SwiftNet.
****
Contrastive Embedding for Generalized Zero-Shot Learning
Zongyan Han, Zhenyong Fu, Shuo Chen, Jian Yang; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2371-2381
Generalized zero-shot learning (GZSL) aims to recognize objects from both seen a
nd unseen classes, when only the labeled examples from seen classes are provided
. Recent feature generation methods learn a generative model that can synthesize
 the missing visual features of unseen classes to mitigate the data-imbalance pr
oblem in GZSL. However, the original visual feature space is suboptimal for GZSL
 classification since it lacks discriminative information. To tackle this issue,
 we propose to integrate the generation model with the embedding model, yielding
 a hybrid GZSL framework. The hybrid GZSL approach maps both the real and the sy
nthetic samples produced by the generation model into an embedding space, where
we perform the final GZSL classification. Specifically, we propose a contrastive
 embedding (CE) for our hybrid GZSL framework. The proposed contrastive embeddin
g can leverage not only the class-wise supervision but also the instance-wise su
pervision, where the latter is usually neglected by existing GZSL researches. We

evaluate our proposed hybrid GZSL framework with contrastive embedding, named CE-GZSL, on five benchmark datasets. The results show that our CEGZSL method can outperform the state-of-the-arts by a significant margin on three datasets. Our codes are available on https://github.com/Hanzy1996/CE-GZSL.
****

Scale-Localized Abstract Reasoning
Yaniv Benny, Niv Pekar, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12557-12565
We consider the abstract relational reasoning task, which is commonly used as an intelligence test. Since some patterns have spatial rationales, while others are only semantic, we propose a multi-scale architecture that processes each query in multiple resolutions. We show that indeed different rules are solved by different resolutions and a combined multi-scale approach outperforms the existing state of the art in this task on all benchmarks by 5-54%. The success of our method is shown to arise from multiple novelties. First, it searches for relational patterns in multiple resolutions, which allows it to readily detect visual relations, such as location, in higher resolution, while allowing the lower resolution module to focus on semantic relations, such as shape type. Second, we optimize the reasoning network of each resolution proportionally to its performance, hereby we motivate each resolution to specialize on the rules for which it performs better than the others and ignore cases that are already solved by the other resolutions. Third, we propose a new way to pool information along the rows and the columns of the illustration-grid of the query. Our work also analyses the existing benchmarks, demonstrating that the RAVEN dataset selects the negative examples in a way that is easily exploited. We, therefore, propose a modified version of the RAVEN dataset, named RAVEN-FAIR. Our code and pretrained models are available at https://github.com/yanivbenny/MRNet.
****

Transferable Query Selection for Active Domain Adaptation
Bo Fu, Zhangjie Cao, Jianmin Wang, Mingsheng Long; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7272-7281
Unsupervised domain adaptation (UDA) enables transferring knowledge from a related source domain to a fully unlabeled target domain. Despite the significant advances in UDA, the performance gap remains quite large between UDA and supervised learning with fully labeled target data. Active domain adaptation (ADA) mitigates the gap under minimal annotation cost by selecting a small quota of target samples to annotate and incorporating them into training. Due to the domain shift, the query selection criteria of prior active learning methods may be ineffective to select the most informative target samples for annotation. In this paper, we propose Transferable Query Selection (TQS), which selects the most informative samples under domain shift by an ensemble of three new criteria: transferable committee, transferable uncertainty, and transferable domainness. We further develop a randomized selection algorithm to enhance the diversity of the selected samples. Experiments show that TQS remarkably outperforms previous UDA and ADA methods on several domain adaptation datasets. Deeper analyses demonstrate that TQS can select the most informative target samples under the domain shift.
****

CLCC: Contrastive Learning for Color Constancy
Yi-Chen Lo, Chia-Che Chang, Hsuan-Chao Chiu, Yu-Hao Huang, Chia-Ping Chen, Yu-Lin Chang, Kevin Jou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8053-8063
In this paper, we present CLCC, a novel contrastive learning framework for color constancy. Contrastive learning has been applied for learning high-quality visual representations for image classification. One key aspect to yield useful representations for image classification is to design illuminant invariant augmentations. However, the illuminant invariant assumption conflicts with the nature of the color constancy task, which aims to estimate the illuminant given a raw image. Therefore, we construct effective contrastive pairs for learning better illuminant-dependent features via a novel raw-domain color augmentation. On the NUS-8 dataset, our method provides 17.5% relative improvements over a strong baseline

, reaching state-of-the-art performance without increasing model complexity. Furthermore, our method achieves competitive performance on the Gehler dataset with 3x fewer parameters compared to top-ranking deep learning methods. More importantly, we show that our model is more robust to different scenes under close proximity of illuminants, significantly reducing 28.7% worst-case error in data-sparse regions. Our code is available at https://github.com/howardyclo/clcc-cvpr21.
****
Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World
Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, Xianglong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8565-8574
Deep learning models are vulnerable to adversarial examples. As a more threatening type for practical deep learning systems, physical adversarial examples have received extensive research attention in recent years. However, without exploiting the intrinsic characteristics such as model-agnostic and human-specific patterns, existing works generate weak adversarial perturbations in the physical world, which fall short of attacking across different models and show visually suspicious appearance. Motivated by the viewpoint that attention reflects the intrinsic characteristics of the recognition process, this paper proposes the Dual Attention Suppression (DAS) attack to generate visually-natural physical adversarial camouflage with strong transferability by suppressing both model and human attention. As for attacking, we generate transferable adversarial camouflages by distracting the model-shared similar attention patterns from the target to non-target regions. Meanwhile, based on the fact that human visual attention always focuses on salient items (e.g., suspicious distortions), we evade the human-specific bottom-up attention to generate visually-natural camouflage which is correlated to the scenario context. We conduct extensive experiments in both the digital and physical world for classification and detection tasks on up to date models (e.g., Yolo-V5) and significantly demonstrate that our method outperforms state-of-the-art methods.
****
Long-Tailed Multi-Label Visual Recognition by Collaborative Training on Uniform and Re-Balanced Samplings
Hao Guo, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15089-15098
Long-tailed data distribution is common in many multi-label visual recognition tasks and the direct use of these data for training usually leads to relatively low performance on tail classes. While re-balanced data sampling can improve the performance on tail classes, it may also hurt the performance on head classes in training due to label co-occurrence. In this paper, we propose a new approach to train on both uniform and re-balanced samplings in a collaborative way, resulting in performance improvement on both head and tail classes. More specifically, we design a visual recognition network with two branches: one takes the uniform sampling as input while the other takes the re-balanced sampling as the input. For each branch, we conduct visual recognition using a binary-cross-entropy-based classification loss with learnable logit compensation. We further define a new cross-branch loss to enforce the consistency when the same input image goes through the two branches. We conduct extensive experiments on VOC-LT and COCO-LT datasets. The results show that the proposed method significantly outperforms previous state-of-the-art methods on long-tailed multi-label visual recognition.
****
3D Object Detection With Pointformer
Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7463-7472
Feature learning for 3D object detection from point clouds is very challenging due to the irregularity of 3D point cloud data. In this paper, we propose Pointformer, a Transformer backbone designed for 3D point clouds to learn features effectively. Specifically, a Local Transformer module is employed to model interacti

ons among points in a local region, which learns context-dependent region features at an object level. A Global Transformer is designed to learn context-aware representations at the scene level. To further capture the dependencies among multi-scale representations, we propose Local-Global Transformer to integrate local features with global features from higher resolution. In addition, we introduce an efficient coordinate refinement module to shift down-sampled points closer to object centroids, which improves object proposal generation. We use Pointformer as the backbone for state-of-the-art object detection models and demonstrate significant improvements over original models on both indoor and outdoor datasets.
****
Fair Feature Distillation for Visual Recognition
Sangwon Jung, Donggyu Lee, Taeeon Park, Taesup Moon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12115-12124
Fairness is becoming an increasingly crucial issue for computer vision, especially in the human-related decision systems. However, achieving algorithmic fairness, which makes a model produce indiscriminative outcomes against protected groups, is still an unresolved problem. In this paper, we devise a systematic approach which reduces algorithmic biases via feature distillation for visual recognition tasks, dubbed as MMD-based Fair Distillation (MFD). While the distillation technique has been widely used in general to improve the prediction accuracy, to the best of our knowledge, there has been no explicit work that also tries to improve fairness via distillation. Furthermore, We give a theoretical justification of our MFD on the effect of knowledge distillation and fairness. Throughout the extensive experiments, we show our MFD significantly mitigates the bias against specific minorities without any loss of the accuracy on both synthetic and real-world face datasets.
****
Diversifying Sample Generation for Accurate Data-Free Quantization
Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, Xianglong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15658-15667
Quantization has emerged as one of the most prevalent approaches to compress and accelerate neural networks. Recently, data-free quantization has been widely studied as a practical and promising solution. It synthesizes data for calibrating the quantized model according to the batch normalization (BN) statistics of FP32 ones and significantly relieves the heavy dependency on real training data in traditional quantization methods. Unfortunately, we find that in practice, the synthetic data identically constrained by BN statistics suffers serious homogenization at both distribution level and sample level and further causes a significant performance drop of the quantized model. We propose Diverse Sample Generation (DSG) scheme to mitigate the adverse effects caused by homogenization. Specifically, we slack the alignment of feature statistics in the BN layer to relax the constraint at the distribution level and design a layerwise enhancement to reinforce specific layers for different data samples. Our DSG scheme is versatile and even able to be applied to the state-of-the-art post-training quantization method like AdaRound. We evaluate the DSG scheme on the large-scale image classification task and consistently obtain significant improvements over various network architectures and quantization methods, especially when quantized to lower bits (e.g., up to 22% improvement on W4A4). Moreover, benefiting from the enhanced diversity, models calibrated by synthetic data perform close to those calibrated by real data and even outperform them on W4A4.
****
SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation
Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, Graham W. Taylor; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5912-5921
In this paper we introduce a Transformer-based approach to video object segmentation (VOS). To address compounding error and scalability issues of prior work, w

e propose a scalable, end-to-end method for VOS called Sparse Spatiotemporal Transformers (SST). SST extracts per-pixel representations for each object in a video using sparse attention over spatiotemporal features. Our attention-based formulation for VOS allows a model to learn to attend over a history of multiple frames and provides suitable inductive bias for performing correspondence-like computations necessary for solving motion segmentation. We demonstrate the effectiveness of attention-based over recurrent networks in the spatiotemporal domain. Our method achieves competitive results on YouTube-VOS and DAVIS 2017 with improved scalability and robustness to occlusions compared with the state of the art. Code is available at https://github.com/dukebw/SSTVOS.
****
Inferring CAD Modeling Sequences Using Zone Graphs
Xianghao Xu, Wenzhe Peng, Chin-Yi Cheng, Karl D.D. Willis, Daniel Ritchie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6062-6070
In computer-aided design (CAD), the ability to "reverse engineer" the modeling steps used to create 3D shapes is a long-sought-after goal. This process can be decomposed into two sub-problems: converting an input mesh or point cloud into a boundary representation (or B-rep), and then inferring modeling operations which construct this B-rep. In this paper, we present a new system for solving the second sub-problem. Central to our approach is a new geometric representation: the zone graph. Zones are the set of solid regions formed by extending all B-Rep faces and partitioning space with them; a zone graph has these zones as its nodes, with edges denoting geometric adjacencies between them. Zone graphs allow us to tractably work with industry-standard CAD operations, unlike prior work using CSG with parametric primitives. We focus on CAD programs consisting of sketch + extrude + Boolean operations, which are common in CAD practice. We phrase our problem as search in the space of such extrusions permitted by the zone graph, and we train a graph neural network to score potential extrusions in order to accelerate the search. We show that our approach outperforms an existing CSG inference baseline in terms of geometric reconstruction accuracy and reconstruction time, while also creating more plausible modeling sequences.
****
Closed-Form Factorization of Latent Semantics in GANs
Yujun Shen, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1532-1540
A rich set of interpretable dimensions has been shown to emerge in the latent space of the Generative Adversarial Networks (GANs) trained for synthesizing images. In order to identify such latent dimensions for image editing, previous methods typically annotate a collection of synthesized samples and train linear classifiers in the latent space. However, they require a clear definition of the target attribute as well as the corresponding manual annotations, limiting their applications in practice. In this work, we examine the internal representation learned by GANs to reveal the underlying variation factors in an unsupervised manner. In particular, we take a closer look into the generation mechanism of GANs and further propose a closed-form factorization algorithm for latent semantic discovery by directly decomposing the pre-trained weights. With a lightning-fast implementation, our approach is capable of not only finding semantically meaningful dimensions comparably to the state-of-the-art supervised methods, but also resulting in far more versatile concepts across multiple GAN models trained on a wide range of datasets.
****
Weakly-Supervised Physically Unconstrained Gaze Estimation
Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9980-9989
A major challenge for physically unconstrained gaze estimation is acquiring training data with 3D gaze annotations for in-the-wild and outdoor scenarios. In contrast, videos of human interactions in unconstrained environments are abundantly available and can be much more easily annotated with frame-level activity label

s. In this work, we tackle the previously unexplored problem of weakly-supervise
d gaze estimation from videos of human interactions. We leverage the insight tha
t strong gaze-related geometric constraints exist when people perform the activi
ty of "looking at each other" (LAEO). To acquire viable 3D gaze supervision from
 LAEO labels, we propose a training algorithm along with several novel loss func
tions especially designed for the task. With weak supervision from two large sca
le CMU-Panoptic and AVA-LAEO activity datasets, we show significant improvements
 in (a) the accuracy of semi-supervised gaze estimation and (b) cross-domain gen
eralization on the state-of-the-art physically unconstrained in-the-wild Gaze360
 gaze estimation benchmark. We open source our code at https://github.com/NVlabs
/weakly-supervised-gaze.
****
A Circular-Structured Representation for Visual Emotion Distribution Learning
Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, Xinbo Gao; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 423
7-4246
Visual Emotion Analysis (VEA) has attracted increasing attention recently with t
he prevalence of sharing images on social networks. Since human emotions are amb
iguous and subjective, it is more reasonable to address VEA in a label distribut
ion learning (LDL) paradigm rather than a single-label classification task. Diff
erent from other LDL tasks, there exist intrinsic relationships between emotions
 and unique characteristics within them, as demonstrated in psychological theori
es. Inspired by this, we propose a well-grounded circular-structured representat
ion to utilize the prior knowledge for visual emotion distribution learning. To
be specific, we first construct an Emotion Circle to unify any emotional state w
ithin it. On the proposed Emotion Circle, each emotion distribution is represent
ed with an emotion vector, which is defined with three attributes (i.e., emotion
 polarity, emotion type, emotion intensity) as well as two properties (i.e., sim
ilarity, additivity). Besides, we design a novel Progressive Circular (PC) loss
to penalize the dissimilarities between predicted emotion vector and labeled one
 in a coarse-to-fine manner, which further boosts the learning process in an emo
tion-specific way. Extensive experiments and comparisons are conducted on public
 visual emotion distribution datasets, and the results demonstrate that the prop
osed method outperforms the state-of-the-art methods.
****
VirTex: Learning Visual Representations From Textual Annotations
Karan Desai, Justin Johnson; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2021, pp. 11162-11173
The de-facto approach to many vision tasks is to start from pretrained visual re
presentations, typically learned via supervised training on ImageNet. Recent met
hods have explored unsupervised pretraining to scale to vast quantities of unlab
eled images. In contrast, we aim to learn high-quality visual representations fr
om fewer images. To this end we revisit supervised pretraining, and seek data-ef
ficient alternatives to classification-based pretraining. We propose VirTex -- a
 pretraining approach using semantically dense captions to learn visual represen
tations. We train convolutional networks from scratch on COCO Captions, and tran
sfer them to downstream recognition tasks including image classification, object
 detection, and instance segmentation. On all tasks, VirTex yields features that
 match or exceed those learned on ImageNet -- supervised or unsupervised -- desp
ite using up to ten times fewer images.
****
MASA-SR: Matching Acceleration and Spatial Adaptation for Reference-Based Image
Super-Resolution
Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, Jiaya Jia; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6368-63
77
Reference-based image super-resolution (RefSR) has shown promising success in re
covering high-frequency details by utilizing an external reference image (Ref).
In this task, texture details are transferred from the Ref image to the low-reso
lution (LR) image according to their point- or patch-wise correspondence. Theref

ore, high-quality correspondence matching is critical. It is also desired to be computationally efficient. Besides, existing RefSR methods tend to ignore the po tential large disparity in distributions between the LR and Ref images, which hu rts the effectiveness of the information utilization. In this paper, we propose the MASA network for RefSR, where two novel modules are designed to address thes e problems. The proposed Match & Extraction Module significantly reduces the com putational cost by a coarse-to-fine correspondence matching scheme. The Spatial Adaptation Module learns the difference of distribution between the LR and Ref i mages, and remaps the distribution of Ref features to that of LR features in a s patially adaptive way. This scheme makes the network robust to handle different reference images. Extensive quantitative and qualitative experiments validate th e effectiveness of our proposed model.
****

Spatiotemporal Contrastive Video Representation Learning
Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belo ngie, Yin Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2021, pp. 6964-6974
We present a self-supervised Contrastive Video Representation Learning (CVRL) me thod to learn spatiotemporal visual representations from unlabeled videos. Our r epresentations are learned using a contrastive loss, where two augmented clips f rom the same short video are pulled together in the embedding space, while clips from different videos are pushed away. We study what makes for good data augmen tations for video self-supervised learning and find that both spatial and tempor al information are crucial. We carefully design data augmentations involving spa tial and temporal cues. Concretely, we propose a temporally consistent spatial a ugmentation method to impose strong spatial augmentations on each frame of the v ideo while maintaining the temporal consistency across frames. We also propose a sampling-based temporal augmentation method to avoid overly enforcing invarianc e on clips that are distant in time. On Kinetics-600, a linear classifier traine d on the representations learned by CVRL achieves 70.4% top-1 accuracy with a 3D -ResNet-50 (R3D-50) backbone, outperforming ImageNet supervised pre-training by 15.7% and SimCLR unsupervised pre-training by 18.8% using the same inflated R3D-50. The performance of CVRL can be further improved to 72.9% with a larger R3D-1 52 (2x filters) backbone, significantly closing the gap between unsupervised and supervised video representation learning. Our code and models will be available at https://github.com/tensorflow/models/tree/master/official/.
****

Scaled-YOLOv4: Scaling Cross Stage Partial Network
Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 130 29-13038
We show that the YOLOv4 object detection neural network based on the CSP approac h, scales both up and down and is applicable to small and large networks while m aintaining optimal speed and accuracy. We propose a network scaling approach tha t modifies not only the depth, width, resolution, but also structure of the netw ork. YOLOv4-large model achieves state-of-the-art results: 55.5% AP (73.4% AP50) for the MS COCO dataset at a speed of 16 FPS on Tesla V100, while with the tes t time augmentation, YOLOv4-large achieves 56.0% AP (73.3 AP50). To the best of our knowledge, this is currently the highest accuracy on the COCO dataset among any published work. The YOLOv4-tiny model achieves 22.0% AP (42.0% AP50) at a sp eed of 443 FPS on RTX 2080Ti, while by using TensorRT, batch size = 4 and FP16-precision the YOLOv4-tiny achieves 1774 FPS.
****

Quantifying Explainers of Graph Neural Networks in Computational Pathology
Guillaume Jaume, Pushpak Pati, Behzad Bozorgtabar, Antonio Foncubierta, Anna Mar ia Anniciello, Florinda Feroce, Tilman Rau, Jean-Philippe Thiran, Maria Gabrani, Orcun Goksel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2021, pp. 8106-8116
Explainability of deep learning methods is imperative to facilitate their clinic al adoption in digital pathology. However, popular deep learning methods and exp

lainability techniques (explainers) based on pixel-wise processing disregard bio logical entities' notion, thus complicating comprehension by pathologists. In th is work, we address this by adopting biological entity-based graph processing an d graph explainers enabling explanations accessible to pathologists. In this con text, a major challenge becomes to discern meaningful explainers, particularly i n a standardized and quantifiable fashion. To this end, we propose herein a set of novel quantitative metrics based on statistics of class separability using pa thologically measurable concepts to characterize graph explainers. We employ the proposed metrics to evaluate three types of graph explainers, namely the layer- wise relevance propagation, gradient-based saliency, and graph pruning approache s, to explain Cell-Graph representations for Breast Cancer Subtyping. The propos ed metrics are also applicable in other domains by using domain-specific intuiti ve concepts. We validate the qualitative and quantitative findings on the BRACS dataset, a large cohort of breast cancer RoIs, by expert pathologists. The code and models will be released upon acceptance.
****
Knowledge Evolution in Neural Networks
Ahmed Taha, Abhinav Shrivastava, Larry S. Davis; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12843-12852
Deep learning relies on the availability of a large corpus of data (labeled or u nlabeled). Thus, one challenging unsettled question is: how to train a deep netw ork on a relatively small dataset? To tackle this question, we propose an evolut ion-inspired training approach to boost performance on relatively small datasets . The knowledge evolution (KE) approach splits a deep network into two hypothese s: the fit-hypothesis and the reset-hypothesis. We iteratively evolve the knowle dge inside the fit-hypothesis by perturbing the reset-hypothesis for multiple ge nerations. This approach not only boosts performance, but also learns a slim net work with a smaller inference cost. KE integrates seamlessly with both vanilla a nd residual convolutional networks. KE reduces both overfitting and the burden f or data collection. We evaluate KE on various network architectures and loss fun ctions. We evaluate KE using relatively small datasets (e.g., CUB-200) and rando mly initialized deep networks. KE achieves an absolute 21% improvement margin on a state-of-the-art baseline. This performance improvement is accompanied by a r elative 73% reduction in inference cost. KE achieves state-of-the-art results on classification and metric learning benchmarks.
****
Revisiting Knowledge Distillation: An Inheritance and Exploration Framework
Zhen Huang, Xu Shen, Jun Xing, Tongliang Liu, Xinmei Tian, Houqiang Li, Bing Den g, Jianqiang Huang, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2021, pp. 3579-3588
Knowledge Distillation (KD) is a popular technique to transfer knowledge from a teacher model or ensemble to a student model. Its success is generally attribute d to the privileged information on similarities/consistency between the class di stributions or intermediate feature representations of the teacher model and the student model. However, directly pushing the student model to mimic the probabi lities/features of the teacher model to a large extent limits the student model in learning undiscovered knowledge/features. In this paper, we propose a novel i nheritance and exploration knowledge distillation framework (IE-KD), in which a student model is split into two parts -- inheritance and exploration. The inheri tance part is learned with a similarity loss to transfer the existing learned kn owledge from the teacher model to the student model, while the exploration part is encouraged to learn representations different from the inherited ones with a dis-similarity loss. Our IE-KD framework is generic and can be easily combined w ith existing distillation or mutual learning methods for training deep neural ne tworks. Extensive experiments demonstrate that these two parts can jointly push the student model to learn more diversified and effective representations, and o ur IE-KD can be a general technique to improve the student network to achieve SO TA performance. Furthermore, by applying our IE-KD to the training of two networ ks, the performance of both can be improved w.r.t. deep mutual learning.
****

Temporally-Weighted Hierarchical Clustering for Unsupervised Action Segmentation
Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11225-11234
Action segmentation refers to inferring boundaries of semantically consistent visual concepts in videos and is an important requirement for many video understanding tasks. For this and other video understanding tasks, supervised approaches have achieved encouraging performance but require a high volume of detailed, frame-level, annotations. We present a fully automatic and unsupervised approach for segmenting actions in a video that does not require any training. Our proposal is an effective temporally-weighted hierarchical clustering algorithm that can group semantically consistent frames of the video. The main finding is that representing a video with a 1-nearest neighbor graph by taking into account the time progression is sufficient to form semantically and temporally consistent clusters of frames where each cluster may represent some action in the video. Additionally, we establish strong unsupervised baselines for action segmentation and show significant performance improvements over published unsupervised methods on five challenging action segmentation datasets. Our code is available at https://github.com/ssarfraz/FINCH-Clustering/tree/master/TW-FINCH
****
SMURF: Self-Teaching Multi-Frame Unsupervised RAFT With Full-Image Warping
Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, Rico Jonschkowski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3887-3896
We present SMURF, a method for unsupervised learning of optical flow that improves state of the art on all benchmarks by 36% to 40% and even outperforms several supervised approaches such as PWC-Net and FlowNet2. Our method integrates architecture improvements from supervised optical flow, i.e. the RAFT model, with new ideas for unsupervised learning that include a novel unsupervised sequence loss and self-supervision loss, a technique for handling out-of-frame motion, and an approach for learning effectively from multi-frame video data while still only requiring two frames for inference.
****
Glancing at the Patch: Anomaly Localization With Global and Local Feature Comparison
Shenzhi Wang, Liwei Wu, Lei Cui, Yujun Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 254-263
Anomaly localization, with the purpose to segment the anomalous regions within images, is challenging due to the large variety of anomaly types. Existing methods typically train deep models by treating the entire image as a whole yet put little effort into learning the local distribution, which is vital for this pixel-precise task. In this work, we propose an unsupervised patch-based approach that gives due consideration to both the global and local information. More concretely, we employ a Local-Net and Global-Net to extract features from any individual patch and its surrounding respectively. Global-Net is trained with the purpose to mimic the local feature such that we can easily detect an abnormal patch when its feature mismatches that from the context. We further introduce an Inconsistency Anomaly Detection (IAD) head and a Distortion Anomaly Detection (DAD) head to sufficiently spot the discrepancy between global and local features. A scoring function derived from the multi-head design facilitates high-precision anomaly localization. Extensive experiments on a couple of real-world datasets suggest that our approach outperforms state-of-the-art competitors by a sufficiently large margin.
****
Single-View 3D Object Reconstruction From Shape Priors in Memory
Shuo Yang, Min Xu, Haozhe Xie, Stuart Perry, Jiahao Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3152-3161
Existing methods for single-view 3D object reconstruction directly learn to transform image features into 3D representations. However, these methods are vulnera

ble to images containing noisy backgrounds and heavy occlusions because the extracted image features do not contain enough information to reconstruct high-quality 3D shapes. Humans routinely use incomplete or noisy visual cues from an image to retrieve similar 3D shapes from their memory and reconstruct the 3D shape of an object. Inspired by this, we propose a novel method, named Mem3D, that explicitly constructs shape priors to supplement the missing information in the image. Specifically, the shape priors are in the forms of "image-voxel" pairs in the memory network, which is stored by a well-designed writing strategy during training. We also propose a voxel triplet loss function that helps to retrieve the precise 3D shapes that are highly related to the input image from shape priors. The LSTM-based shape encoder is introduced to extract information from the retrieved 3D shapes, which are useful in recovering the 3D shape of an object that is heavily occluded or in complex environments. Experimental results demonstrate that Mem3D significantly improves reconstruction quality and performs favorably against state-of-the-art methods on the ShapeNet and Pix3D datasets.
****
Recognizing Actions in Videos From Unseen Viewpoints
AJ Piergiovanni, Michael S. Ryoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4124-4132
Standard methods for video recognition use large CNNs designed to capture spatio-temporal data. However, training these models requires a large amount of labeled training data, containing a wide variety of actions, scenes, settings and camera viewpoints. In this paper, we show that current convolutional neural network models are unable to recognize actions from camera viewpoints not present in their training data (i.e., unseen view action recognition). To address this, we develop approaches based on 3D pose and introduce a new geometric convolutional layer that can learn viewpoint invariant representations. Further, we introduce a new, challenging dataset for unseen view recognition and show the approaches ability to learn viewpoint invariant representations.
****
Perceptual Indistinguishability-Net (PI-Net): Facial Image Obfuscation With Manipulable Semantics
Jia-Wei Chen, Li-Ju Chen, Chia-Mu Yu, Chun-Shien Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6478-6487
With the growing use of camera devices, the industry has many image datasets that provide more opportunities for collaboration between the machine learning community and industry. However, the sensitive information in the datasets discourages data owners from releasing these datasets. Despite recent research devoted to removing sensitive information from images, they provide neither meaningful privacy-utility trade-off nor provable privacy guarantees. In this study, with the consideration of the perceptual similarity, we propose perceptual indistinguishability (PI) as a formal privacy notion particularly for images. We also propose PI-Net, a privacy-preserving mechanism that achieves image obfuscation with PI guarantee. Our study shows that PI-Net achieves significantly better privacy utility trade-off through public image data.
****
To the Point: Efficient 3D Object Detection in the Range Image With Graph Convolution Kernels
Yuning Chai, Pei Sun, Jiquan Ngiam, Weiyue Wang, Benjamin Caine, Vijay Vasudevan, Xiao Zhang, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16000-16009
3D object detection is vital for many robotics applications. For tasks where a 2D perspective range image exists, we propose to learn a 3D representation directly from this range image view. To this end, we designed a 2D convolutional network architecture that carries the 3D spherical coordinates of each pixel throughout the network. Its layers can consume any arbitrary convolution kernel in place of the default inner product kernel and exploit the underlying local geometry around each pixel. We outline four such kernels: a dense kernel according to the bag-of-words paradigm, and three graph kernels inspired by recent graph neural n

etwork advances: the Transformer, the PointNet, and the Edge Convolution. We also explore cross-modality fusion with the camera image, facilitated by operating in the perspective range image view. Our method performs competitively on the Waymo Open Dataset and improves the state-of-the-art AP for pedestrian detection from 69.7% to 75.5%. It is also efficient in that our smallest model, which still outperforms the popular PointPillars in quality, requires 180 times fewer FLOPS and model parameters.
****
Coarse-To-Fine Domain Adaptive Semantic Segmentation With Photometric Alignment and Category-Center Regularization
Haoyu Ma, Xiangru Lin, Zifeng Wu, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4051-4060
Unsupervised domain adaptation (UDA) in semantic segmentation is a fundamental yet promising task relieving the need for laborious annotation works. However, the domain shifts/discrepancies problem in this task compromise the final segmentation performance. Based on our observation, the main causes of the domain shifts are differences in imaging conditions, called image-level domain shifts, and differences in object category configurations called category-level domain shifts. In this paper, we propose a novel UDA pipeline that unifies image-level alignment and category-level feature distribution regularization in a coarse-to-fine manner. Specifically, on the coarse side, we propose a photometric alignment module that aligns an image in the source domain with a reference image from the target domain using a set of image-level operators; on the fine side, we propose a category-oriented triplet loss that imposes a soft constraint to regularize category centers in the source domain and a self-supervised consistency regularization method in the target domain. Experimental results show that our proposed pipeline improves the generalization capability of the final segmentation model and significantly outperforms all previous state-of-the-arts.
****
Self-Supervised Wasserstein Pseudo-Labeling for Semi-Supervised Image Classification
Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12267-12277
The goal is to use Wasserstein metric to provide pseudo labels for the unlabeled images to train a Convolutional Neural Networks (CNN) in a Semi-Supervised Learning (SSL) manner for the classification task. The basic premise in our method is that the discrepancy between two discrete empirical measures (e.g., clusters) which come from the same or similar distribution is expected to be less than the case where these measures come from completely two different distributions. In our proposed method, we first pre-train our CNN using a self-supervised learning method to make a cluster assumption on the unlabeled images. Next, inspired by the Wasserstein metric which considers the geometry of the metric space to provide a natural notion of similarity between discrete empirical measures, we leverage it to cluster the unlabeled images and then match the clusters to their similar class of labeled images to provide a pseudo label for the data within each cluster. We have evaluated and compared our method with state-of-the-art SSL methods on the standard datasets to demonstrate its effectiveness.
****
MeanShift++: Extremely Fast Mode-Seeking With Applications to Segmentation and Object Tracking
Jennifer Jang, Heinrich Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4102-4113
MeanShift is a popular mode-seeking clustering algorithm used in a wide range of applications in machine learning. However, it is known to be prohibitively slow, with quadratic runtime per iteration. We propose MeanShift++, an extremely fast mode-seeking algorithm based on MeanShift that uses a grid-based approach to speed up the mean shift step, replacing the computationally expensive neighbors search with a density-weighted mean of adjacent grid cells. In addition, we show that this grid-based technique for density estimation comes with theoretical gua

rantees. The runtime is linear in the number of points and exponential in dimens ion, which makes MeanShift++ ideal on low-dimensional applications such as image segmentation and object tracking. We provide extensive experimental analysis sh owing that MeanShift++ can be more than 10,000x faster than MeanShift with compe titive clustering results on benchmark datasets and nearly identical image segme ntations as MeanShift. Finally, we show promising results for object tracking.
****
PCLs: Geometry-Aware Neural Reconstruction of 3D Pose With Perspective Crop Laye rs
Frank Yu, Mathieu Salzmann, Pascal Fua, Helge Rhodin; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9064-9 073
Local processing is an essential feature of CNNs and other neural network archit ectures -- it is one of the reasons why they work so well on images where releva nt information is, to a large extent, local. However, perspective effects stemmi ng from the projection in a conventional camera vary for different global positi ons in the image. We introduce Perspective Crop Layers (PCLs) -- a form of persp ective crop of the region of interest based on the camera geometry -- and show t hat accounting for the perspective consistently improves the accuracy of state-o f-the-art 3D pose reconstruction methods. PCLs are modular neural network layers , which, when inserted into existing CNN and MLP architectures, deterministicall y remove the location-dependent perspective effects while leaving end-to-end tra ining and the number of parameters of the underlying neural network unchanged. W e demonstrate that PCL leads to improved 3D human pose reconstruction accuracy f or CNN architectures that use cropping operations, such as spatial transformer n etworks (STN), and, somewhat surprisingly, MLPs used for 2D-to-3D keypoint lifti ng. Our conclusion is that it is important to utilize camera calibration informa tion when available, for classical and deep-learning-based computer vision alike . PCL offers an easy way to improve the accuracy of existing 3D reconstruction n etworks by making them geometry aware. Our code is publicly available at github. com/yu-frank/PerspectiveCropLayers.
****
Partially View-Aligned Representation Learning With Noise-Robust Contrastive Los s
Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 21, pp. 1134-1143
In real-world applications, it is common that only a portion of data is aligned across views due to spatial, temporal, or spatiotemporal asynchronism, thus lead ing to the so-called Partially View-aligned Problem (PVP). To solve such a less- touched problem without the help of labels, we propose simultaneously learning r epresentation and aligning data using a noise-robust contrastive loss. In brief, for each sample from one view, our method aims to identify its within-category counterparts from other views, and thus the cross-view correspondence could be e stablished. As the contrastive learning needs data pairs as input, we construct positive pairs using the known correspondences and negative pairs using random s ampling. To alleviate or even eliminate the influence of the false negatives cau sed by random sampling, we propose a noise-robust contrastive loss that could ad aptively prevent the false negatives from dominating the network optimization. T o the best of our knowledge, this could be the first successful attempt of enabl ing contrastive learning robust to noisy labels. In fact, this work might remark ably enrich the learning paradigm with noisy labels. More specifically, the trad itional noisy labels are defined as incorrect annotations for the supervised tas ks such as classification. In contrast, this work proposes that the view corresp ondence might be false, which is remarkably different from the widely-accepted d efinition of noisy label. Extensive experiments show the promising performance o f our method comparing with 10 state-of-the-art multi-view approaches in the clu stering and classification tasks. The code will be publicly released at https:// pengxi.me.
****

i3DMM: Deep Implicit 3D Morphable Model of Human Heads
Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12803-12813
We present the first deep implicit 3D morphable model (i3DMM) of full heads. Unlike earlier morphable face models it not only captures identity-specific geometry, texture, and expressions of the frontal face, but also models the entire head, including hair. We collect a new dataset consisting of 64 people with different expressions and hairstyles to train i3DMM. Our approach has the following favorable properties: (i) It is the first full head morphable model that includes hair. (ii) In contrast to mesh-based models it can be trained on merely rigidly aligned scans, without requiring difficult non-rigid registration. (iii) We design a novel architecture to decouple the shape model into an implicit reference shape and a deformation of this reference shape. With that, dense correspondences between shapes can be learned implicitly. (iv) This architecture allows us to semantically disentangle the geometry and color components, as color is learned in the reference space. Geometry is further disentangled as identity, expressions, and hairstyle, while color is disentangled as identity and hairstyle components. We show the merits of i3DMM using ablation studies, comparisons to state-of-the-art models, and applications such as semantic head editing and texture transfer. We will make our model publicly available.
****
Searching by Generating: Flexible and Efficient One-Shot NAS With Architecture Generator
Sian-Yao Huang, Wei-Ta Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 983-992
In one-shot NAS, sub-networks need to be searched from the supernet to meet different hardware constraints. However, the search cost is high and N times of searches are needed for N different constraints. In this work, we propose a novel search strategy called architecture generator to search sub-networks by generating them, so that the search process can be much more efficient and flexible. With the trained architecture generator, given target hardware constraints as the input, N good architectures can be generated for N constraints by just one forward pass without re-searching and supernet retraining. Moreover, we propose a novel single-path supernet, called unified supernet, to further improve search efficiency and reduce GPU memory consumption of the architecture generator. With the architecture generator and the unified supernet, we propose a flexible and efficient one-shot NAS framework, called Searching by Generating NAS (SGNAS). With the pre-trained supernt, the search time of SGNAS for N different hardware constraints is only 5 GPU hours, which is 4N times faster than previous SOTA single-path methods. After training from scratch, the top1-accuracy of SGNAS on ImageNet is 77.1%, which is comparable with the SOTAs. The code is available at: https://github.com/eric8607242/SGNAS.
****
Discovering Interpretable Latent Space Directions of GANs Beyond Binary Attributes
Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12177-12185
Generative adversarial networks (GANs) learn to map noise latent vectors to high-fidelity image outputs. It is found that the input latent space shows semantic correlations with the output image space. Recent works aim to interpret the latent space and discover meaningful directions that correspond to human interpretable image transformations. However, these methods either rely on explicit scores of attributes (e.g., memorability) or are restricted to binary ones (e.g., gender), which largely limits the applicability of editing tasks, especially for free-form artistic tasks like style/anime editing. In this paper, we propose an adversarial method, AdvStyle, for discovering interpretable directions in the absence of well-labeled scores or binary attributes. In particular, the proposed adversarial method simultaneously optimizes the discovered directions and the attribu

te assessor using the target attribute data as positive samples, while the gener ated ones being negative. In this way, arbitrary attributes can be edited by col lecting positive data only, and the proposed method learns a controllable repres entation enabling manipulation of non-binary attributes like anime styles and fa cial characteristics. Moreover, the proposed learning strategy attenuates the en tanglement between attributes, such that multiattribute manipulation can be easi ly achieved without any additional constraint. Furthermore, we reveal several in teresting semantics with the involuntarily learned negative directions. Extensiv e experiments on 9 anime attributes and 7 human attributes demonstrate the effec tiveness of our adversarial approach qualitatively and quantitatively. Code is a vailable at https://github.com/BERYLSHEEP/AdvStyle.
****

ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis
Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, J ing Shao, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2021, pp. 4360-4369
The rapid progress of photorealistic synthesis techniques has reached at a criti cal point where the boundary between real and manipulated images starts to blur. Thus, benchmarking and advancing digital forgery analysis have become a pressin g issue. However, existing face forgery datasets either have limited diversity o r only support coarse-grained analysis. To counter this emerging threat, we cons truct the ForgeryNet dataset, an extremely large face forgery dataset with unifi ed annotations in image- and video-level data across four tasks: 1) Image Forger y Classification, including two-way (real / fake), three-way (real / fake with i dentity-replaced forgery approaches / fake with identity-remained forgery approa ches), and n-way (real and 15 respective forgery approaches) classification. 2) Spatial Forgery Localization, which segments the manipulated area of fake images compared to their corresponding source real images. 3) Video Forgery Classifica tion, which re-defines the video-level forgery classification with manipulated f rames in random positions. This task is important because attackers in real worl d are free to manipulate any target frame. and 4) Temporal Forgery Localization, to localize the temporal segments which are manipulated. ForgeryNet is by far t he largest publicly available deep face forgery dataset in terms of data-scale (2.9 million images, 221,247 videos), manipulations (7 image-level approaches, 8 video-level approaches), perturbations (36 independent and more mixed perturbati ons) and annotations (6.3 million classification labels, 2.9 million manipulated area annotations and 221,247 temporal forgery segment labels). We perform exten sive benchmarking and studies of existing face forensics methods and obtain seve ral valuable observations.We hope that the scale, quality, and variety of Forger yNet dataset will foster further research and innovation in the area of face for gery classification, spatial and temporal forgery localization etc.
****

Blocks-World Cameras
Jongho Lee, Mohit Gupta; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2021, pp. 11412-11422
For several vision and robotics applications, 3D geometry of man-made environmen ts such as indoor scenes can be represented with a small number of dominant plan es. However, conventional 3D vision techniques typically first acquire dense 3D point clouds before estimating the compact piece-wise planar representations (e. g., by plane-fitting). This approach is costly, both in terms of acquisition and computational requirements, and potentially unreliable due to noisy point cloud s. We propose Blocks-World Cameras, a class of imaging systems which directly re cover dominant planes of piece-wise planar scenes (Blocks-World), without requir ing point clouds. The Blocks-World Cameras are based on a structured-light syste m projecting a single pattern with a sparse set of cross-shaped features. We dev elop a novel geometric algorithm for recovering scene planes without explicit co rrespondence matching, thereby avoiding computationally intensive search or opti mization routines. The proposed approach has low device and computational comple xity, and requires capturing only one or two images. We demonstrate highly effic ient and precise planar-scene sensing with simulations and real experiments, acr

oss various imaging conditions, including defocus blur, large lighting variation
s, ambient illumination, and scene clutter.
****
The Affective Growth of Computer Vision
Norman Makoto Su, David J. Crandall; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2021, pp. 9291-9300
The success of deep learning has led to intense growth and interest in computer
vision, along with concerns about its potential impact on society. Yet we know l
ittle about how these changes have affected the people that research and practic
e computer vision: we as a community spend so much effort trying to replicate th
e abilities of humans, but so little time considering the impact of this work on
 ourselves. In this paper, we report on a study in which we asked computer visio
n researchers and practitioners to write stories about emotionally-salient event
s that happened to them. Our analysis of over 50 responses found tremendous affe
ctive (emotional) strain in the computer vision community. While many describe e
xcitement and success, we found strikingly frequent feelings of isolation, cynic
ism, apathy, and exasperation over the state of the field. This is especially tr
ue among people who do not share the unbridled enthusiasm for normative standard
s for computer vision research and who do not see themselves as part of the "in-
crowd." Our findings suggest that these feelings are closely tied to the kinds o
f research and professional practices now expected in computer vision. We argue
that as a community with significant stature, we need to work towards an inclusi
ve culture that makes transparent and addresses the real emotional toil of its m
embers.
****
Lifelong Person Re-Identification via Adaptive Knowledge Accumulation
Nan Pu, Wei Chen, Yu Liu, Erwin M. Bakker, Michael S. Lew; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7
901-7910
Person ReID methods always learn through a stationary domain that is fixed by th
e choice of a given dataset. In many contexts (e.g., lifelong learning), those m
ethods are ineffective because the domain is continually changing in which case
incremental learning over multiple domains is required potentially. In this work
 we explore a new and challenging ReID task, namely lifelong person re-identific
ation (LReID), which enables to learn continuously across multiple domains and e
ven generalise on new and unseen domains. Following the cognitive processes in t
he human brain, we design an Adaptive Knowledge Accumulation (AKA) framework tha
t is endowed with two crucial abilities: knowledge representation and knowledge
operation. Our method alleviates catastrophic forgetting on seen domains and dem
onstrates the ability to generalize to unseen domains. Correspondingly, we also
provide a new and large-scale benchmark for LReID. Extensive experiments demonst
rate our method outperforms other competitors by a margin of 5.8% mAP in general
ising evaluation.
****
Omnimatte: Associating Objects and Their Effects in Video
Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T. Freeman, Mich
ael Rubinstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2021, pp. 4507-4515
Computer vision has become increasingly better at segmenting objects in images a
nd videos; however, scene effects related to the objects -- shadows, reflections
, generated smoke, etc. -- are typically overlooked. Identifying such scene effe
cts and associating them with the objects producing them is important for improv
ing our fundamental understanding of visual scenes, and applications such as rem
oving, duplicating, or enhancing objects in video. We take a step towards solvin
g this novel problem of automatically associating objects with their effects in
video. Given an ordinary video and a rough segmentation mask over time of one or
 more subjects of interest, we estimate an omnimatte for each subject -- an alph
a matte and color image that includes the subject along with all its related tim
e-varying scene elements. Our model is trained only on the input video in a self
-supervised manner, without any manual labels, and is generic -- it produces omn

imattes automatically for arbitrary objects and a variety of effects. We show results on real-world videos containing interactions between different types of subjects (cars, animals, people) and complex effects, ranging from semi-transparent smoke and reflections to fully opaque objects attached to the subject.
****
Detecting Human-Object Interaction via Fabricated Compositional Learning
Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14646-14655
Human-Object Interaction (HOI) detection, inferring the relationships between human and objects from images/videos, is a fundamental task for high-level scene understanding. However, HOI detection usually suffers from the open long-tailed nature of interactions with objects, while human has extremely powerful compositional perception ability to cognize rare or unseen HOI samples. Inspired by this, we devise a novel HOI compositional learning framework, termed as Fabricated Compositional Learning (FCL), to address the problem of open long-tailed HOI detection. Specifically, we introduce an object fabricator to generate effective object representations, and then combine verbs and fabricated objects to compose new HOI samples. With the proposed object fabricator, we are able to generate large-scale HOI samples for rare and unseen categories to alleviate the open long-tailed issues in HOI detection. Extensive experiments on the most popular HOI detection dataset, HICO-DET, demonstrate the effectiveness of the proposed method for imbalanced HOI detection and significantly improve the state-of-the-art performance on rare and unseen HOI categories.
****
Memory-Efficient Network for Large-Scale Video Compressive Sensing
Ziheng Cheng, Bo Chen, Guanliang Liu, Hao Zhang, Ruiying Lu, Zhengjue Wang, Xin Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16246-16255
Video snapshot compressive imaging (SCI) captures a sequence of video frames in a single shot using a 2D detector. The underlying principle is that during one exposure time, different masks are imposed on the high-speed scene to form a compressed measurement. With the knowledge of masks, optimization algorithms or deep learning methods are employed to reconstruct the desired high-speed video frames from this snapshot measurement. Unfortunately, though these methods can achieve decent results, the long running time of optimization algorithms or huge training memory occupation of deep networks still preclude them in practical applications. In this paper, we develop a memory-efficient network for large-scale video SCI based on multi-group reversible 3D convolutional neural networks. In addition to the basic model for the grayscale SCI system, we take one step further to combine demosaicing and SCI reconstruction to directly recover color video from Bayer measurements. Extensive results on both simulation and real data captured by SCI cameras demonstrate that our proposed model outperforms previous state-of-the-art with less memory and thus can be used in large-scale problems. The code is at https: //github.com/BoChenGroup/RevSCI-net.
****
Deep Optimized Priors for 3D Shape Modeling and Reconstruction
Mingyue Yang, Yuxin Wen, Weikai Chen, Yongwei Chen, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3269-3278
Many learning-based approaches have difficulty scaling to unseen data, as the generality of its learned prior is limited to the scale and variations of the training samples. This holds particularly true with 3D learning tasks, given the sparsity of 3D datasets available. We introduce a new learning framework for 3D modeling and reconstruction that greatly improves the generalization ability of a deep generator. Our approach strives to connect the good ends of both learning-based and optimization-based methods. In particular, unlike the common practice that fixes the pre-trained priors at test time, we propose to further optimize the learned prior and latent code according to the input physical measurements after the training. We show that the proposed strategy effectively breaks the barrie

rs constrained by the pre-trained priors and could lead to high-quality adaptati
on to unseen data. We realize our framework using the implicit surface represent
ation and validate the efficacy of our approach in a variety of challenging task
s that take highly sparse or collapsed observations as input. Experimental resul
ts show that our approach compares favorably with the state-of-the-art methods i
n terms of both generality and accuracy.
****
Affordance Transfer Learning for Human-Object Interaction Detection
Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, Dacheng Tao; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
495-504
Reasoning the human-object interactions (HOI) is essential for deeper scene unde
rstanding, while object affordances (or functionalities) are of great importance
 for human to discover unseen HOIs with novel objects. Inspired by this, we intr
oduce an affordance transfer learning approach to jointly detect HOIs with novel
 object and recognize affordances. Specifically, HOI representations can be deco
upled into a combination of affordance and object representations, making it pos
sible to compose novel interactions by combining affordance representations and
novel object representations from additional images, i.e. transferring the affor
dance to novel objects. With the proposed affordance transfer learning, the mode
l is also capable of inferring the affordances of novel objects from known affor
dance representations. The proposed method can thus be used to 1) improve the pe
rformance of HOI detection, especially for the HOIs with unseen objects; and 2)
infer the affordances of novel objects. Experimental results on two datasets, HI
CO-DET and HOI-COCO (from V-COCO), demonstrate significant improvements over rec
ent state-of-the-art methods for HOI detection and object affordance detection.
Code is available at https://github.com/zhihou7/HOI-CL.
****
DSC-PoseNet: Learning 6DoF Object Pose Estimation via Dual-Scale Consistency
Zongxin Yang, Xin Yu, Yi Yang; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2021, pp. 3907-3916
Compared to 2D object bounding-box labeling, it is very difficult for humans to
annotate 3D object poses, especially when depth images of scenes are unavailable
. This paper investigates whether we can estimate the object poses effectively w
hen only RGB images and 2D object annotations are given. To this end, we present
 a two-step pose estimation framework to attain 6DoF object poses from 2D object
 bounding-boxes. In the first step, the framework learns to segment objects from
 real and synthetic data in a weakly-supervised fashion, and the segmentation ma
sks will act as a prior for pose estimation. In the second step, we design a dua
l-scale pose estimation network, namely DSC-PoseNet, to predict object poses by
employing a differential renderer. To be specific, our DSC-PoseNet firstly predi
cts object poses in the original image scale by comparing the segmentation masks
 and the rendered visible object masks. Then, we resize object regions to a fixe
d scale to estimate poses once again. In this fashion, we eliminate large scale
variations and focus on rotation estimation, thus facilitating pose estimation.
Moreover, we exploit the initial pose estimation to generate pseudo ground-truth
 to train our DSC-PoseNet in a self-supervised manner. The estimation results in
 these two scales are ensembled as our final pose estimation. Extensive experime
nts on widely-used benchmarks demonstrate that our method outperforms state-of-t
he-art models trained on synthetic data by a large margin and even is on par wit
h several fully-supervised methods.
****
Rethinking Graph Neural Architecture Search From Message-Passing
Shaofei Cai, Liang Li, Jincan Deng, Beichen Zhang, Zheng-Jun Zha, Li Su, Qingmin
g Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R
ecognition (CVPR), 2021, pp. 6657-6666
Graph neural networks (GNNs) emerged recently as a standard toolkit for learning
 from data on graphs. Current GNN designing works depend on immense human expert
ise to explore different message-passing mechanisms, and require manual enumerat
ion to determine the proper message-passing depth. Inspired by the strong search

ing capability of neural architecture search (NAS) in CNN, this paper proposes Graph Neural Architecture Search (GNAS) with novel-designed search space. The GNAS can automatically learn better architecture with the optimal depth of message passing on the graph. Specifically, we design Graph Neural Architecture Paradigm (GAP) with tree-topology computation procedure and two types of fine-grained atomic operations (feature filtering & neighbor aggregation) from message-passing mechanism to construct powerful graph network search space. Feature filtering performs adaptive feature selection, and neighbor aggregation captures structural information and calculates neighbors' statistics. Experiments show that our GNAS can search for better GNNs with multiple message-passing mechanisms and optimal message-passing depth. The searched network achieves remarkable improvement over state-of-the-art manual designed and search-based GNNs on five large-scale datasets at three classical graph tasks.
****
Locate Then Segment: A Strong Pipeline for Referring Image Segmentation
Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9858-9867
Referring image segmentation aims to segment the objects referred by a natural language expression. Previous methods usually focus on designing an implicit and recurrent feature interaction mechanism to fuse the visual-linguistic features to directly generate the final segmentation mask without explicitly modeling the localization of the referent guided by language expression and designing a powerful segmentation module. To tackle these problems, we view this task from another perspective by decoupling it into a "locate-then-segment" (LTS) scheme. Given a language expression, people generally first perform attention to the corresponding target image regions, then generate a segmentation mask about the object based on its context. The LTS first extracts and fuses both visual and textual features to get a cross-modal representation, then applies a cross-model interaction on the visual-textual features to locate the referred object with position prior, and finally generates the segmentation result with a light-weight network. Our LTS is simple but surprisingly effective. On three popular benchmark datasets, the LTS outperforms all the previous state-of-the-arts methods by a large margin (e.g., +3.2% on RefCOCO+ and +3.4% on RefCOCOg). In addition, our model is more interpretable with explicitly locating the object, which is also proved by visualization experiments. Accordingly, this framework is very promising to serve as a pipeline for referring image segmentation.
****
Exploring Complementary Strengths of Invariant and Equivariant Representations for Few-Shot Learning
Mamshad Nayeem Rizve, Salman Khan, Fahad Shahbaz Khan, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10836-10846
In many real-world problems, collecting a large number of labeled samples is infeasible. Few-shot learning (FSL) is the dominant approach to address this issue, where the objective is to quickly adapt to novel categories in presence of a limited number of samples. FSL tasks have been predominantly solved by leveraging the ideas from gradient-based meta-learning and metric learning approaches. However, recent works have demonstrated the significance of powerful feature representations with a simple embedding network that can outperform existing sophisticated FSL algorithms. In this work, we build on this insight and propose a novel training mechanism that simultaneously enforces equivariance and invariance to a general set of geometric transformations. Equivariance or invariance has been employed standalone in the previous works; however, to the best of our knowledge, they have not been used jointly. Simultaneous optimization for both of these contrasting objectives allows the model to jointly learn features that are not only independent of the input transformation but also the features that encode the structure of geometric transformations. These complementary sets of features help generalize well to novel classes with only a few data samples. We achieve additional improvements by incorporating a novel self-supervised distillation objecti

ve. Our extensive experimentation shows that even without knowledge distillation our proposed method can outperform current state-of-the-art FSL methods on five popular benchmark datasets.
****
Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation
Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2287-2296
We present a generic image-to-image translation framework, pixel2style2pixel (pSp). Our pSp framework is based on a novel encoder network that directly generates a series of style vectors which are fed into a pretrained StyleGAN generator, forming the extended W+ latent space. We first show that our encoder can directly embed real images into W+, with no additional optimization. Next, we propose utilizing our encoder to directly solve image-to-image translation tasks, defining them as encoding problems from some input domain into the latent domain. By deviating from the standard invert first, edit later methodology used with previous StyleGAN encoders, our approach can handle a variety of tasks even when the input image is not represented in the StyleGAN domain. We show that solving translation tasks through StyleGAN significantly simplifies the training process, as no adversary is required, has better support for solving tasks without pixel-to-pixel correspondence, and inherently supports multi-modal synthesis via the resampling of styles. Finally, we demonstrate the potential of our framework on a variety of facial image-to-image translation tasks, even when compared to state-of-the-art solutions designed specifically for a single task, and further show that it can be extended beyond the human facial domain. Code is available at https://github.com/eladrich/pixel2style2pixel.
****
Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning
Shaoxiang Chen, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8425-8435
Dense Event Captioning (DEC) aims to jointly localize and describe multiple events of interest in untrimmed videos, which is an advancement of the conventional video captioning task (generating a single sentence description for a trimmed video). Weakly Supervised Dense Event Captioning (WS-DEC) goes one step further by not relying on human-annotated temporal event boundaries. However, there are few methods trying to tackle this task, and how to connect localization and description remains an open problem. In this paper, we demonstrate that under weak supervision, the event captioning module and localization module should be more closely bridged in order to improve description performance. Different from previous approaches, in our method, the event captioner generates a sentence from a video segment and feeds it to the sentence localizer to reconstruct the segment, and the localizer produces word importance weights as a guidance for the captioner to improve event description. To further bridge the sentence localizer and event captioner, a concept learner is adopted as the basis of the sentence localizer, which can be utilized to construct an induced set of concept features to enhance video features and improve the event captioner. Finally, our proposed method outperforms state-of-the-art WS-DEC methods on the ActivityNet Captions dataset.
****
DER: Dynamically Expandable Representation for Class Incremental Learning
Shipeng Yan, Jiangwei Xie, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3014-3023
We address the problem of class incremental learning, which is a core step towards achieving adaptive vision intelligence. In particular, we consider the task setting of incremental learning with limited memory and aim to achieve a better stability-plasticity trade-off. To this end, we propose a novel two-stage learning approach that utilizes a dynamically expandable representation for more effective incremental concept modeling. Specifically, at each incremental step, we freeze the previously learned representation and augment it with additional feature dimensions from a new learnable feature extractor. Moreover, we dynamically exp

and the representation according to the complexity of novel concepts by introducing a channel-level mask-based pruning strategy. This enables us to integrate new visual concepts with retaining learned knowledge. Furthermore, we introduce an auxiliary loss to encourage the model to learn diverse and discriminate features for novel concepts. We conduct extensive experiments on the three class incremental learning benchmarks and our method consistently outperforms other methods with a large margin.
****
Fine-Grained Angular Contrastive Learning With Coarse Labels
Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, Leonid Karlinsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8730-8740
Few-shot learning methods offer pre-training techniques optimized for easier later adaptation of the model to new classes (unseen during training) using one or a few examples. This adaptivity to unseen classes is especially important for many practical applications where the pre-trained label space cannot remain fixed for effective use and the model needs to be "specialized" to support new categories on the fly. One particularly interesting scenario, essentially overlooked by the few-shot literature, is Coarse-to-Fine Few-Shot (C2FS), where the training classes (e.g. animals) are of much `coarser granularity' than the target (test) classes (e.g. breeds). A very practical example of C2FS is when the target classes are sub-classes of the training classes. Intuitively, it is especially challenging as (both regular and few-shot) supervised pre-training tends to learn to ignore intra-class variability which is essential for separating sub-classes. In this paper, we introduce a novel 'Angular normalization' module that allows to effectively combine supervised and self-supervised contrastive pre-training to approach the proposed C2FS task, demonstrating significant gains in a broad study over multiple baselines and datasets. We hope that this work will help to pave the way for future research on this new, challenging, and very practical topic of C2FS classification.
****
Polarimetric Normal Stereo
Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 682-690
We introduce a novel method for recovering per-pixel surface normals from a pair of polarization cameras. Unlike past methods that use polarimetric observations as auxiliary features for correspondence matching, we fully integrate them in cost volume construction and filtering to directly recover per-pixel surface normals, not as byproducts of recovered disparities. Our key idea is to introduce a polarimetric cost volume of distance defined on the polarimetric observations and the polarization state computed from the surface normal. We adapt a belief propagation algorithm to filter this cost volume. The filtering algorithm simultaneously estimates the disparities and surface normals as separate entities, while effectively denoising the original noisy polarimetric observations of a quad-Bayer polarization camera. In addition, in contrast to past methods, we model polarimetric light reflection of mesoscopic surface roughness, which is essential to account for its illumination-dependency. We demonstrate the effectiveness of our method on a number of complex, real objects. Our method offers a simple and detailed 3D sensing capability for complex, non-Lambertian surfaces.
****
Manifold Regularized Dynamic Network Pruning
Yehui Tang, Yunhe Wang, Yixing Xu, Yiping Deng, Chao Xu, Dacheng Tao, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5018-5028
Neural network pruning is an essential approach for reducing the computational complexity of deep models so that they can be well deployed on resource-limited devices. Compared with conventional methods, the recently developed dynamic pruning methods determine redundant filters variant to each input instance which achieves higher acceleration. Most of the existing methods discover effective sub-ne

tworks for each instance independently and do not utilize the relationship betwe
en different inputs. To maximally excavate redundancy in the given network archi
tecture, this paper proposes a new paradigm that dynamically removes redundant f
ilters by embedding the manifold information of all instances into the space of
pruned networks (dubbed as ManiDP). We first investigate the recognition complex
ity and feature similarity between images in the training set. Then, the manifol
d relationship between instances and the pruned sub-networks will be aligned in
the training procedure. The effectiveness of the proposed method is verified on
several benchmarks, which shows better performance in terms of both accuracy and
 computational cost compared to the state-of-the-art methods. For example, our m
ethod can reduce 55.3% FLOPs of ResNet-34 with only 0.57% top-1 accuracy degrada
tion on ImageNet.
****
ViPNAS: Efficient Video Pose Estimation via Neural Architecture Search
Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang,
 Xiaogang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2021, pp. 16072-16081
Human pose estimation has achieved significant progress in recent years. However
, most of the recent methods focus on improving accuracy using complicated model
s and ignoring real-time efficiency. To achieve a better trade-off between accur
acy and efficiency, we propose a novel neural architecture search (NAS) method,
termed ViPNAS, to search networks in both spatial and temporal levels for fast o
nline video pose estimation. In the spatial level, we carefully design the searc
h space with five different dimensions including network depth, width, kernel si
ze, group number, and attentions. In the temporal level, we search from a series
 of temporal feature fusions to optimize the total accuracy and speed across mul
tiple video frames. To the best of our knowledge, we are the first to search for
 the temporal feature fusion and automatic computation allocation in videos. Ext
ensive experiments demonstrate the effectiveness of our approach on the challeng
ing COCO2017 and PoseTrack2018 datasets. Our discovered model family, S-ViPNAS a
nd T-ViPNAS, achieve significantly higher inference speed (CPU real-time) withou
t sacrificing the accuracy compared to the previous state-of-the-art methods.
****
Open Domain Generalization with Domain-Augmented Meta-Learning
Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, Mingsheng Long; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 9624-9633
Leveraging datasets available to learn a model with high generalization ability
to unseen domains is important for computer vision, especially when the unseen d
omain's annotated data are unavailable. We study the problem of learning from di
fferent source domains to achieve high performance on an unknown target domain,
where the distributions and label sets of each individual source domain and the
target domain are different. The problem can be generally applied to diverse sou
rce domains and widely applicable to real-world applications. We propose a Domai
n-Augmented Meta-Learning framework to learn open-domain generalizable represent
ations. We augment domains on both feature-level by a new Dirichlet mixup and la
bel-level by distilled soft-labeling, which complements each domain with missing
 classes and other domain knowledge. We conduct meta-learning over domains by de
signing new meta-learning tasks and losses to preserve domain unique knowledge a
nd generalize knowledge across domains simultaneously. Experiment results on var
ious multi-domain datasets demonstrate that the proposed Domain-Augmented Meta-L
earning outperforms previous methods for unseen target classification.
****
DeepTag: An Unsupervised Deep Learning Method for Motion Tracking on Cardiac Tag
ging Magnetic Resonance Images
Meng Ye, Mikael Kanski, Dong Yang, Qi Chang, Zhennan Yan, Qiaoying Huang, Leon A
xel, Dimitris Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2021, pp. 7261-7271
Cardiac tagging magnetic resonance imaging (t-MRI) is the gold standard for regi
onal myocardium deformation and cardiac strain estimation. However, this techniq

ue has not been widely used in clinical diagnosis, as a result of the difficulty of motion tracking encountered with t-MRI images. In this paper, we propose a novel deep learning-based fully unsupervised method for in vivo motion tracking on t-MRI images. We first estimate the motion field (INF) between any two consecutive t-MRI frames by a bi-directional generative diffeomorphic registration neural network. Using this result, we then estimate the Lagrangian motion field between the reference frame and any other frame through a differentiable composition layer. By utilizing temporal information to perform reasonable estimations on spatio-temporal motion fields, this novel method provides a useful solution for motion tracking and image registration in dynamic medical imaging. Our method has been validated on a representative clinical t-MRI dataset; the experimental results show that our method is superior to conventional motion tracking methods in terms of landmark tracking accuracy and inference efficiency.
****
Learning by Planning: Language-Guided Global Image Editing
Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13590-13599
Recently, language-guided global image editing draws increasing attention with growing application potentials. However, previous GAN-based methods are not only confined to domain-specific, low-resolution data but also lacking in interpretability. To overcome the collective difficulties, we develop a text-to-operation model to map the vague editing language request into a series of editing operations, e.g., change contrast, brightness, and saturation. Each operation is interpretable and differentiable. Furthermore, the only supervision in the task is the target image, which is insufficient for a stable training of sequential decisions. Hence, we propose a novel operation planning algorithm to generate possible editing sequences from the target image as pseudo ground truth. Comparison experiments on the newly collected MA5k-Req dataset and GIER dataset show the advantages of our methods. Code is available at https://github.com/jshi31/T2ONet.
****
Curriculum Graph Co-Teaching for Multi-Target Domain Adaptation
Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5351-5360
In this paper we address multi-target domain adaptation (MTDA), where given one labeled source dataset and multiple unlabeled target datasets that differ in data distributions, the task is to learn a robust predictor for all the target domains. We identify two key aspects that can help to alleviate multiple domain-shifts in the MTDA: feature aggregation and curriculum learning. To this end, we propose Curriculum Graph Co-Teaching (CGCT) that uses a dual classifier head, with one of them being a graph convolutional network (GCN) which aggregates features from similar samples across the domains. To prevent the classifiers from over-fitting on its own noisy pseudo-labels we develop a co-teaching strategy with the dual classifier head that is assisted by curriculum learning to obtain more reliable pseudo-labels. Furthermore, when the domain labels are available, we propose Domain-aware Curriculum Learning (DCL), a sequential adaptation strategy that first adapts on the easier target domains, followed by the harder ones. We experimentally demonstrate the effectiveness of our proposed frameworks on several benchmarks and advance the state-of-the-art in the MTDA by large margins (e.g. +5.6% on the DomainNet).
****
Uncalibrated Neural Inverse Rendering for Photometric Stereo of General Surfaces
Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3804-3814
This paper presents an uncalibrated deep neural network framework for the photometric stereo problem. For training models to solve the problem, existing neural network-based methods either require exact light directions or ground-truth surface normals of the object or both. However, in practice, it is challenging to pr

ocure both of this information precisely, which restricts the broader adoption of photometric stereo algorithms for vision application. To bypass this difficulty, we propose an uncalibrated neural inverse rendering approach to this problem. Our method first estimates the light directions from the input images and then optimizes an image reconstruction loss to calculate the surface normals, bidirectional reflectance distribution function value, and depth. Additionally, our formulation explicitly models the concave and convex parts of a complex surface to consider the effects of interreflections in the image formation process. Extensive evaluation of the proposed method on the challenging subjects generally shows comparable or better results than the supervised and classical approaches.
****
Improving the Transferability of Adversarial Samples With Adversarial Transformations
Weibin Wu, Yuxin Su, Michael R. Lyu, Irwin King; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9024-9033
Although deep neural networks (DNNs) have achieved tremendous performance in diverse vision challenges, they are surprisingly susceptible to adversarial examples, which are born of intentionally perturbing benign samples in a human-imperceptible fashion. It thus poses security concerns on the deployment of DNNs in practice, particularly in safety- and security-sensitive domains. To investigate the robustness of DNNs, transfer-based attacks have attracted a growing interest recently due to their high practical applicability, where attackers craft adversarial samples with local models and employ the resultant samples to attack a remote black-box model. However, existing transfer-based attacks frequently suffer from low success rates due to overfitting to the adopted local model. To boost the transferability of adversarial samples, we propose to improve the robustness of synthesized adversarial samples via adversarial transformations. Specifically, we employ an adversarial transformation network to model the most harmful distortions that can destroy adversarial noises and require the synthesized adversarial samples to become resistant to such adversarial transformations. Extensive experiments on the ImageNet benchmark showcase the superiority of our method to state-of-the-art baselines in attacking both undefended and defended models.
****
Self-Supervised Learning for Semi-Supervised Temporal Action Proposal
Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1905-1914
Self-supervised learning presents a remarkable performance to utilize unlabeled data for various video tasks. In this paper, we focus on applying the power of self-supervised methods to improve semi-supervised action proposal generation. Particularly, we design a Self-supervised Semi-supervised Temporal Action Proposal (SSTAP) framework. The SSTAP contains two crucial branches, i.e., temporal-aware semi-supervised branch and relation-aware self-supervised branch. The semi-supervised branch improves the proposal model by introducing two temporal perturbations, i.e., temporal feature shift and temporal feature flip, in the mean teacher framework. The self-supervised branch defines two pretext tasks, including masked feature reconstruction and clip-order prediction, to learn the relation of temporal clues. By this means, SSTAP can better explore unlabeled videos, and improve the discriminative abilities of learned action features. We extensively evaluate the proposed SSTAP on THUMOS14 and ActivityNet v1.3 datasets. The experimental results demonstrate that SSTAP significantly outperforms state-of-the-art semi-supervised methods and even matches fully-supervised methods. The code will be released once this paper is accepted.
****
Learning Compositional Representation for 4D Captures With Neural ODE
Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5340-5350
Learning based representation has become the key to the success of many computer vision systems. While many 3D representations have been proposed, it is still a

n unaddressed problem how to represent a dynamically changing 3D object. In this paper, we introduce a compositional representation for 4D captures, i.e. a deforming 3D object over a temporal span, that disentangles shape, initial state, and motion respectively. Each component is represented by a latent code via a trained encoder. To model the motion, a neural Ordinary Differential Equation (ODE) is trained to update the initial state conditioned on the learned motion code, and a decoder takes the shape code and the updated state code to reconstruct the 3D model at each time stamp. To this end, we propose an Identity Exchange Training (IET) strategy to encourage the network to learn effectively decoupling each component. Extensive experiments demonstrate that the proposed method outperforms existing state-of-the-art deep learning based methods on 4D reconstruction, and significantly improves on various tasks, including motion transfer and completion.
****
Effective Snapshot Compressive-Spectral Imaging via Deep Denoising and Total Variation Priors
Haiquan Qiu, Yao Wang, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9127-9136
Snapshot compressive imaging (SCI) is a new type of compressive imaging system that compresses multiple frames of images into a single snapshot measurement, which enjoys low cost, low bandwidth, and high-speed sensing rate. By applying the existing SCI methods to deal with hyperspectral images, however, could not fully exploit the underlying structures, and thereby demonstrate unsatisfactory reconstruction performance. To remedy such issue, this paper aims to propose a new effective method by taking advantage of two intrinsic priors of the hyperspectral images, namely deep image denoising and total variation (TV) priors. Specifically, we propose an optimization objective to utilize these two priors. By solving this optimization objective, our method is equivalent to incorporate a weighted FFDNet and a 2DTV or 3DTV denoiser into the plug-and-play framework. Extensive numerical experiments demonstrate the outperformance of the proposed method over several state-of-the-art alternatives. Additionally, we provide a detailed convergence analysis of the resulting plug-and-play algorithm under relatively weak conditions such as without using diminishing step sizes. The code is available at https://github.com/ucker/SCI-TV-FFDNet.
****
LAFEAT: Piercing Through Adversarial Defenses With Latent Features
Yunrui Yu, Xitong Gao, Cheng-Zhong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5735-5745
Deep convolutional neural networks are susceptible to adversarial attacks. They can be easily deceived to give an incorrect output by adding a tiny perturbation to the input. This presents a great challenge in making CNNs robust against such attacks. An influx of new defense techniques have been proposed to this end. In this paper, we show that latent features in certain "robust" models are surprisingly susceptible to adversarial attacks. On top of this, we introduce a unified Linfinity white-box attack algorithm which harnesses latent features in its gradient descent steps, namely LAFEAT. We show that not only is it computationally much more efficient for successful attacks, but it is also a stronger adversary than the current state-of-the-art across a wide range of defense mechanisms. This suggests that model robustness could be contingent the effective use of the defender's hidden components, and it should no longer be viewed from a holistic perspective.
****
Exploiting Spatial Dimensions of Latent in GAN for Real-Time Image Editing
Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, Youngjung Uh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 852-861
Generative adversarial networks (GANs) synthesize realistic images from random latent vectors. Although manipulating the latent vectors controls the synthesized outputs, editing real images with GANs suffers from i) time-consuming optimization for projecting real images to the latent vectors, ii) or inaccurate embeddin

g through an encoder. We propose StyleMapGAN: the intermediate latent space has spatial dimensions, and a spatially variant modulation replaces AdaIN. It makes the embedding through an encoder more accurate than existing optimization-based methods while maintaining the properties of GANs. Experimental results demonstrate that our method significantly outperforms state-of-the-art models in various image manipulation tasks such as local editing and image interpolation. Last but not least, conventional editing methods on GANs are still valid on our StyleMapGAN. Source code is available at https://github.com/naver-ai/StyleMapGAN.
****
Bidirectional Projection Network for Cross Dimension Scene Understanding
Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, Tien-Tsin Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14373-14382
2D image representations are in regular grids and can be processed efficiently, whereas 3D point clouds are unordered and scattered in 3D space. The information inside these two visual domains is well complementary, e.g., 2D images have fine-grained texture while 3D point clouds contain plentiful geometry information. However, most current visual recognition systems process them individually. In this paper, we present a bidirectional projection network (BPNet) for joint 2D and 3D reasoning in an end-to-end manner. It contains 2D and 3D sub-networks with symmetric architectures, that are connected by our proposed bidirectional projection module (BPM). Via the BPM, complementary 2D and 3D information can interact with each other in multiple architectural levels, such that advantages in these two visual domains can be combined for better scene recognition. Extensive quantitative and qualitative experimental evaluations show that joint reasoning over 2D and 3D visual domains can benefit both 2D and 3D scene understanding simultaneously. Our BPNet achieves top performance on the ScanNetV2 benchmark for both 2D and 3D semantic segmentation. Code is available at https://github.com/wbhu/BPNet.
****
Event-Based Synthetic Aperture Imaging With a Hybrid Network
Xiang Zhang, Wei Liao, Lei Yu, Wen Yang, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14235-14244
Synthetic aperture imaging (SAI) is able to achieve the see through effect by blurring out the off-focus foreground occlusions and reconstructing the in-focus occluded targets from multi-view images. However, very dense occlusions and extreme lighting conditions may bring significant disturbances to the SAI based on conventional frame-based cameras, leading to performance degeneration. To address these problems, we propose a novel SAI system based on the event camera which can produce asynchronous events with extremely low latency and high dynamic range. Thus, it can eliminate the interference of dense occlusions by measuring with almost continuous views, and simultaneously tackle the over/under exposure problems. To reconstruct the occluded targets, we propose a hybrid encoder-decoder network composed of spiking neural networks (SNNs) and convolutional neural networks (CNNs). In the hybrid network, the spatio-temporal information of the collected events is first encoded by SNN layers, and then transformed to the visual image of the occluded targets by a style-transfer CNN decoder. Through experiments, the proposed method shows remarkable performance in dealing with very dense occlusions and extreme lighting conditions, and high quality visual images can be reconstructed using pure event data.
****
RSG: A Simple but Effective Module for Learning Imbalanced Datasets
Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, Zhenghua Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3784-3793
Imbalanced datasets widely exist in practice and are a great challenge for training deep neural models with a good generalization on infrequent classes. In this work, we propose a new rare-class sample generator (RSG) to solve this problem. RSG aims to generate some new samples for rare classes during training, and it

has in particular the following advantages: (1) it is convenient to use and high
ly versatile, because it can be easily integrated into any kind of convolutional
 neural network, and it works well when combined with different loss functions,
and (2) it is only used during the training phase, and therefore, no additional
burden is imposed on deep neural networks during the testing phase. In extensive
 experimental evaluations, we verify the effectiveness of RSG. Furthermore, by l
everaging RSG, we obtain competitive results on Imbalanced CIFAR and new state-o
f-the-art results on Places-LT, ImageNet-LT, and iNaturalist 2018. The source co
de is available at https://github.com/Jianf-Wang/RSG.
****
Learning Statistical Texture for Semantic Segmentation
Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, Junjie Yan; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021
, pp. 12537-12546
Existing semantic segmentation works mainly focus on learning the contextual inf
ormation in high-level semantic features with CNNs. In order to maintain a preci
se boundary, low-level texture features are directly skip-connected into the dee
per layers. Nevertheless, texture features are not only about local structure, b
ut also include global statistical knowledge of the input image. In this paper,
we fully take advantages of the low-level texture features and propose a novel S
tatistical Texture Learning Network (STLNet) for semantic segmentation. For the
first time, STLNet analyzes the distribution of low level information and effici
ently utilizes them for the task. Specifically, a novel Quantization and Countin
g Operator (QCO) is designed to describe the texture information in a statistica
l manner. Based on QCO, two modules are introduced: (1) Texture Enhance Module (
TEM), to capture texture-related information and enhance the texture details; (2
) Pyramid Texture Feature Extraction Module (PTFEM), to effectively extract the
statistical texture features from multiple scales. Through extensive experiments
, we show that the proposed STLNet achieves state-of-the-art performance on thre
e semantic segmentation benchmarks: Cityscapes, PASCAL Context and ADE20K.
****
Neural Feature Search for RGB-Infrared Person Re-Identification
Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, Zongyuan Sun; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
pp. 587-597
RGB-Infrared person re-identification (RGB-IR ReID) is a challenging cross-modal
ity retrieval problem, which aims at matching the person-of-interest over visibl
e and infrared camera views. Most existing works achieve performance gains throu
gh manually-designed feature selection modules, which often require significant
domain knowledge and rich experience. In this paper, we study a general paradigm
, termed Neural Feature Search (NFS), to automate the process of feature selecti
on. Specifically, NFS combines a dual-level feature search space and a different
iable search strategy to jointly select identity-related cues in coarse-grained
channels and fine-grained spatial pixels. This combination allows NFS to adaptiv
ely filter background noises and concentrate on informative parts of human bodie
s in a data-driven manner. Moreover, a cross-modality contrastive optimization s
cheme further guides NFS to search features that can minimize modality discrepan
cy whilst maximizing inter-class distance. Extensive experiments on mainstream b
enchmarks demonstrate that our method outperforms state-of-the-arts, especially
achieving better performance on the RegDB dataset with significant improvement o
f 11.20% and 8.64% in Rank-1 and mAP, respectively.
****
FP-NAS: Fast Probabilistic Neural Architecture Search
Zhicheng Yan, Xiaoliang Dai, Peizhao Zhang, Yuandong Tian, Bichen Wu, Matt Feisz
li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2021, pp. 15139-15148
Differential Neural Architecture Search (NAS) requires all layer choices to be h
eld in memory simultaneously; this limits the size of both search space and fina
l architecture. In contrast, Probabilistic NAS, such as PARSEC, learns a distrib
ution over high-performing architectures, and uses only as much memory as needed

to train a single model. Nevertheless, it needs to sample many architectures, making it computationally expensive for searching in an extensive space. To solve these problems, we propose a sampling method adaptive to the distribution entropy, drawing more samples to encourage explorations at the beginning, and reducing samples as learning proceeds. Furthermore, to search fast in the multi-variate space, we propose a coarse-to-fine strategy by using a factorized distribution at the beginning which can reduce the number of architecture parameters by over an order of magnitude. We call this method Fast Probabilistic NAS (FP-NAS). Compared with PARSEC, it can sample 64% fewer architectures and search 2.1x faster. Compared with FBNetV2, FP-NAS is 1.9x - 3.5x faster, and the searched models outperform FBNetV2 models on ImageNet. FP-NAS allows us to expand the giant FBNetV2 space to be wider (i.e. larger channel choices) and deeper (i.e. more blocks), while adding Split-Attention block and enabling the search over the number of splits. When searching a model of size 0.4G FLOPS, FP-NAS is 132x faster than EfficientNet, and the searched FP-NAS-L0 model outperforms EfficientNet-B0 by 0.7% accuracy. Without using any architecture surrogate or scaling tricks, we directly search large models up to 1.0G FLOPS. Our FP-NAS-L2 model with simple distillation outperforms BigNAS-XL with advanced in-place distillation by 0.7% accuracy using similar FLOPS.
****
Fast Sinkhorn Filters: Using Matrix Scaling for Non-Rigid Shape Correspondence With Functional Maps
Gautam Pai, Jing Ren, Simone Melzi, Peter Wonka, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 384-393
In this paper, we provide a theoretical foundation for pointwise map recovery from functional maps and highlight its relation to a range of shape correspondence methods based on spectral alignment. With this analysis in hand, we develop a novel spectral registration technique: Fast Sinkhorn Filters, which allows for the recovery of accurate and bijective pointwise correspondences with a superior time and memory complexity in comparison to existing approaches. Our method combines the simple and concise representation of correspondence using functional maps with the matrix scaling schemes from computational optimal transport. By exploiting the sparse structure of the kernel matrices involved in the transport map computation, we provide an efficient trade-off between acceptable accuracy and complexity for the problem of dense shape correspondence.
****
Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction
Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10472-10481
This paper considers a new problem of adapting a pre-trained model of human mesh reconstruction to out-of-domain streaming videos. However, most previous methods based on the parametric SMPL model underperform in new domains with unexpected, domain-specific attributes, such as camera parameters, lengths of bones, backgrounds, and occlusions. Our general idea is to dynamically fine-tune the source model on test video streams with additional temporal constraints, such that it can mitigate the domain gaps without over-fitting the 2D information of individual test frames. A subsequent challenge is how to avoid conflicts between the 2D and temporal constraints. We propose to tackle this problem using a new training algorithm named Bilevel Online Adaptation (BOA), which divides the optimization process of overall multi-objective into two steps of weight probe and weight update in a training iteration. We demonstrate that BOA leads to state-of-the-art results on two human mesh reconstruction benchmarks.
****
The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth
Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, Michael Firman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1164-1174
Self-supervised monocular depth estimation networks are trained to predict scene

 depth using nearby frames as a supervision signal during training. However, for
 many applications, sequence information in the form of video frames is also ava
ilable at test time. The vast majority of monocular networks do not make use of
this extra signal, thus ignoring valuable information that could be used to impr
ove the predicted depth. Those that do, either use computationally expensive tes
t-time refinement techniques or off-the-shelf recurrent networks, which only ind
irectly make use of the geometric information that is inherently available. We p
ropose ManyDepth, an adaptive approach to dense depth estimation that can make u
se of sequence information at test time, when it is available. Taking inspiratio
n from multi-view stereo, we propose a deep end-to-end cost volume based approac
h that is trained using self-supervision only. We present a novel consistency lo
ss that encourages the network to ignore the cost volume when it is deemed unrel
iable, e.g. in the case of moving objects, and an augmentation scheme to cope wi
th static cameras. Our detailed experiments on both KITTI and Cityscapes show th
at we outperform all published self-supervised baselines, including those that u
se single or multiple frames at test time.
****
Distribution-Aware Adaptive Multi-Bit Quantization
Sijie Zhao, Tao Yue, Xuemei Hu; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 9281-9290
In this paper, we explore the compression of deep neural networks by quantizing
the weights and activations into multi-bit binary networks (MBNs). A distributio
n-aware multi-bit quantization (DMBQ) method that incorporates the distribution
prior into the optimization of quantization is proposed. Instead of solving the
optimization in each iteration, DMBQ search the optimal quantization scheme over
 the distribution space beforehand, and select the quantization scheme during tr
aining using a fast lookup table based strategy. Based upon DMBQ, we further pro
pose loss-guided bit-width allocation (LBA) to adaptively quantize and even prun
e the neural network. The first-order Taylor expansion is applied to build a met
ric for evaluating the loss sensitivity of the quantization of each channel, and
 automatically adjust the bit-width of weights and activations channel-wisely. W
e extend our method to image classification tasks and experimental results show
that our method not only outperforms state-of-the-art quantized networks in term
s of accuracy but also is more efficient in terms of training time compared with
 state-of-the-art MBNs, even for the extremely low bit width (below 1-bit) quant
ization cases.
****
KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Bas
ed VQA
Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, Marcus Rohrbach; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2021, pp. 14111-14121
One of the most challenging question types in VQA is when answering the question
 requires outside knowledge not present in the image. In this work we study open
-domain knowledge, the setting when the knowledge required to answer a question
is not given/annotated, neither at training nor test time. We tap into two types
 of knowledge representations and reasoning. First, implicit knowledge which can
 be learned effectively from unsupervised language pretraining and supervised tr
aining data with transformer-based models. Second, explicit, symbolic knowledge
encoded in knowledge bases. Our approach combines both---exploiting the powerful
 implicit reasoning of transformer models for answer prediction, and integrating
 symbolic representations from a knowledge graph, while never losing their expli
cit semantics to an implicit embedding. We combine diverse sources of knowledge
to cover the wide variety of knowledge needed to solve knowledge-based questions
. We show our approach, KRISP, significantly outperforms state-of-the-art on OK-
VQA, the largest available dataset for open-domain knowledge-based VQA. We show
with extensive ablations that while our model successfully exploits implicit kno
wledge reasoning, the symbolic answer module which explicitly connects the knowl
edge graph to the answer vocabulary is critical to the performance of our method
 and generalizes to rare answers.

****
Amalgamating Knowledge From Heterogeneous Graph Neural Networks
Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15709-15718
In this paper, we study a novel knowledge transfer task in the domain of graph neural networks (GNNs). We strive to train a multi-talented student GNN, without accessing human annotations, that "amalgamates" knowledge from a couple of teacher GNNs with heterogeneous architectures and handling distinct tasks. The student derived in this way is expected to integrate the expertise from both teachers while maintaining a compact architecture. To this end, we propose an innovative approach to train a slimmable GNN that enables learning from teachers with varying feature dimensions. Meanwhile, to explicitly align topological semantics between the student and teachers, we introduce a topological attribution map (TAM) to highlight the structural saliency in a graph, based on which the student imitates the teachers' ways of aggregating information from neighbors. Experiments on seven datasets across various tasks, including multi-label classification and joint segmentation-classification, demonstrate that the learned student, with a lightweight architecture, achieves gratifying results on par with and sometimes even superior to those of the teachers in their specializations. Our code is publicly available at https://github.com/ycjing/AmalgamateGNN.PyTorch.
****
MetaSets: Meta-Learning on Point Sets for Generalizable Representations
Chao Huang, Zhangjie Cao, Yunbo Wang, Jianmin Wang, Mingsheng Long; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8863-8872
Deep learning techniques for point clouds have achieved strong performance on a range of 3D vision tasks. However, it is costly to annotate large-scale point sets, making it critical to learn generalizable representations that can transfer well across different point sets. In this paper, we study a new problem of 3D Domain Generalization (3DDG) with the goal to generalize the model to other unseen domains of point clouds without any access to them in the training process. It is a challenging problem due to the substantial geometry shift from simulated to real data, such that most existing 3D models underperform due to overfitting the complete geometries in the source domain. We propose to tackle this problem with MetaSets, which meta-learns point cloud representations from a set of classification tasks on carefully-designed transformed point sets containing specific geometry priors. The learned representations are more generalizable to various unseen domains of different geometries. We design two benchmarks for Sim-to-Real transfer of 3D point clouds. Experimental results show that MetaSets outperforms existing 3D deep learning methods by large margins.
****
StEP: Style-Based Encoder Pre-Training for Multi-Modal Image Synthesis
Moustafa Meshry, Yixuan Ren, Larry S. Davis, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3712-3721
We propose a novel approach for multi-modal Image-to-image (I2I) translation. To tackle the one-to-many relationship between input and output domains, previous works use complex training objectives to learn a latent embedding, jointly with the generator, that models the variability of the output domain. In contrast, we directly model the style variability of images, independent of the image synthesis task. Specifically, we pre-train a generic style encoder using a novel proxy task to learn an embedding of images, from arbitrary domains, into a low-dimensional style latent space. The learned latent space introduces several advantages over previous traditional approaches to multi-modal I2I translation. First, it is not dependent on the target dataset, and generalizes well across multiple domains. Second, it learns a more powerful and expressive latent space, which improves the fidelity of style capture and transfer. The proposed style pre-training also simplifies the training objective and speeds up the training significantly. Furthermore, we provide a detailed study of the contribution of different loss

terms to the task of multi-modal I2I translation, and propose a simple alternative to VAEs to enable sampling from unconstrained latent spaces. Finally, we achieve state-of-the-art results on six challenging benchmarks with a simple training objective that includes only a GAN loss and a reconstruction loss.
****
Goal-Oriented Gaze Estimation for Zero-Shot Learning
Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, Tatsuya Harada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3794-3803
Zero-shot learning (ZSL) aims to recognize novel classes by transferring semantic knowledge from seen classes to unseen classes. Since semantic knowledge is built on attributes shared between different classes, which are highly local, strong prior for localization of object attribute is beneficial for visual-semantic embedding. Interestingly, when recognizing unseen images, human would also automatically gaze at regions with certain semantic clue. Therefore, we introduce a novel goal-oriented gaze estimation module (GEM) to improve the discriminative attribute localization based on the class-level attributes for ZSL. We aim to predict the actual human gaze location to get the visual attention regions for recognizing a novel object guided by attribute description. Specifically, the task-dependent attention is learned with the goal-oriented GEM, and the global image features are simultaneously optimized with the regression of local attribute features. Experiments on three ZSL benchmarks, i.e., CUB, SUN and AWA2, show the superiority or competitiveness of our proposed method against the state-of-the-art ZSL methods. The ablation analysis on real gaze data CUB-VWSW also validates the benefits and accuracy of our gaze estimation module. This work implies the promising benefits of collecting human gaze dataset and automatic gaze estimation algorithms on high-level computer vision tasks.
****
LED2-Net: Monocular 360deg Layout Estimation via Differentiable Depth Rendering
Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, Yi-Hsuan Tsai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12956-12965
Although significant progress has been made in room layout estimation, most methods aim to reduce the loss in the 2D pixel coordinate rather than exploiting the room structure in the 3D space. Towards reconstructing the room layout in 3D, we formulate the task of 360 layout estimation as a problem of predicting depth on the horizon line of a panorama. Specifically, we propose the Differentiable Depth Rendering procedure to make the conversion from layout to depth prediction differentiable, thus making our proposed model end-to-end trainable while leveraging the 3D geometric information, without the need of providing the ground truth depth. Our method achieves state-of-the-art performance on numerous 360 layout benchmark datasets. Moreover, our formulation enables a pre-training step on the depth dataset, which further improves the generalizability of our layout estimation model.
****
Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos
Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12669-12678
We address the problem of localizing a specific moment from an untrimmed video by a language sentence query. Generally, previous methods mainly exist two problems that are not fully solved: 1) How to effectively model the fine-grained visual-language alignment between video and language query? 2) How to accurately localize the moment in the original video length? In this paper, we streamline the temporal language localization as a novel multi-stage aggregated transformer network. Specifically, we first introduce a new visual-language transformer backbone, which enables iterations and alignments among all elements in visual and language sequences. Different from previous multi-modal transformers, our backbone keeps both structure unified and modality specific. Moreover, we also propose a mu

lti-stage aggregation module topped on the transformer backbone. In this module, we compute three stage-specific representations corresponding to different moment stages respectively, i.e. starting, middle and ending stages, for each video element. Then for a moment candidate, we concatenate the starting/middle/ending representations of its starting/middle/ending elements respectively to form the final moment representation. Because the obtained moment representation captures the stage specific information, it is very discriminative for accurate localization. Extensive experiments on ActivityNet Captions and TACoS datasets demonstrate our proposed method achieves significant improvements compared with all other methods.
****
DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation
Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, Song Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15769-15778
Semantic segmentation of nighttime images plays an equally important role as that of daytime images in autonomous driving, but the former is much more challenging due to poor illuminations and arduous human annotations. In this paper, we propose a novel domain adaptation network (DANNet) for nighttime semantic segmentation without using labeled nighttime image data. It employs an adversarial training with a labeled daytime dataset and an unlabeled dataset that contains coarsely aligned day-night image pairs. Specifically, for the unlabeled day-night image pairs, we use the pixel-level predictions of static object categories on a daytime image as a pseudo supervision to segment its counterpart nighttime image. We further design a re-weighting strategy to handle the inaccuracy caused by misalignment between day-night image pairs and wrong predictions of daytime images, as well as boost the prediction accuracy of small objects. The proposed DANNet is the first one stage adaptation framework for nighttime semantic segmentation, which does not train additional day-night image transfer models as a separate pre-processing stage. Extensive experiments on Dark Zurich and Nighttime Driving datasets show that our method achieves state-of-the-art performance for nighttime semantic segmentation.
****
Dynamic Transfer for Multi-Source Domain Adaptation
Yunsheng Li, Lu Yuan, Yinpeng Chen, Pei Wang, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10998-11007
Recent works of multi-source domain adaptation focus on learning a domain-agnostic model, of which the parameters are static. However, such a static model is difficult to handle conflicts across multiple domains, and suffers from a performance degradation in both source domains and target domain. In this paper, we present dynamic transfer to address domain conflicts, where the model parameters are adapted to samples. The key insight is that adapting model across domains is achieved via adapting model across samples. Thus, it breaks down source domain barriers and turns multi-source domains into a single source domain. This also simplifies the alignment between source and target domains, as it only requires the target domain to be aligned with any part of the union of source domains. Furthermore, we find dynamic transfer can be simply modeled by aggregating residual matrices and a static convolution matrix. Experimental results show that, without using domain labels, our dynamic transfer outperforms the state-of-the-art method by more than 3% on the large multi-source domain adaptation datasets -- DomainNet.
****
Semi-Supervised Video Deraining With Dynamical Rain Generator
Zongsheng Yue, Jianwen Xie, Qian Zhao, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 642-652
While deep learning (DL)-based video deraining methods have achieved significant success recently, they still exist two major drawbacks. Firstly, most of them do not sufficiently model the characteristics of rain layers of rainy videos. In

fact, the rain layers exhibit strong physical properties (e.g., direction, scale and thickness) in spatial dimension and natural continuities in temporal dimension, and thus can be generally modelled by the spatial-temporal process in statistics. Secondly, current DL-based methods seriously depend on the labeled synthetic training data, whose rain types are always deviated from those in unlabeled real data. Such gap between synthetic and real data sets leads to poor performance when applying them in real scenarios. Against these issues, this paper proposes a new semisupervised video deraining method, in which a dynamic rain generator is employed to fit the rain layer, expecting to better depict its insightful characteristics. Specifically, such dynamic generator consists of one emission model and one transition model to simultaneously encode the spatially physical structure and temporally continuous changes of rain streaks, respectively, which both are parameterized as deep neural networks (DNNs). Further more, different prior formats are designed for the labeled synthetic and unlabeled real data, so as to fully exploit the common knowledge underlying them. Last but not least, we also design a Monte Carlo EM algorithm to solve this model. Extensive experiments are conducted to verify the superiorities of the proposed semi-supervised deraining model.
****
See Through Gradients: Image Batch Recovery via GradInversion
Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16337-16346
Training deep neural networks requires gradient estimation from data batches to update parameters. Gradients per parameter are averaged over a set of data and this has been presumed to be safe for privacy-preserving training in joint, collaborative, and federated learning applications. Prior work only showed the possibility of recovering input data given gradients under very restrictive conditions - a single input point, or a network with no non-linearities, or a small 32x32 px input batch. Therefore, averaging gradients over larger batches was thought to be safe. In this work, we introduce GradInversion, using which input images from a larger batch (8 - 48 images) can also be recovered for large networks such as ResNets (50 layers), on complex datasets such as ImageNet (1000 classes, 224x224 px). We formulate an optimization task that converts random noise into natural images, matching gradients while regularizing image fidelity. We also propose an algorithm for target class label recovery given gradients. We further propose a group consistency regularization framework, where multiple agents starting from different random seeds work together to find an enhanced reconstruction of original data batch. We show that gradients encode a surprisingly large amount of information, such that all the individual images can be recovered with high fidelity via GradInversion, even for complex datasets, deep networks, and large batch sizes.
****
Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition
Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, Hanzi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7660-7669
In this paper, we propose a novel Feature Decomposition and Reconstruction Learning (FDRL) method for effective facial expression recognition. We view the expression information as the combination of the shared information (expression similarities) across different expressions and the unique information (expression-specific variations) for each expression. More specifically, FDRL mainly consists of two crucial networks: a Feature Decomposition Network (FDN) and a Feature Reconstruction Network (FRN). In particular, FDN first decomposes the basic features extracted from a backbone network into a set of facial action-aware latent features to model expression similarities. Then, FRN captures the intra-feature and inter-feature relationships for latent features to characterize expression-specific variations, and reconstructs the expression feature. To this end, two modules including an intra-feature relation modeling module and an inter-feature relat

ion modeling module are developed in FRN. Experimental results on both the in-the-lab databases (including CK+, MMI, and Oulu-CASIA) and the in-the-wild databases (including RAF-DB and SFEW) show that the proposed FDRL method consistently achieves higher recognition accuracy than several state-of-the-art methods. This clearly highlights the benefit of feature decomposition and reconstruction for classifying expressions.
****
Seeing Behind Objects for 3D Multi-Object Tracking in RGB-D Sequences
Norman Muller, Yu-Shiang Wong, Niloy J. Mitra, Angela Dai, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6071-6080
Multi-object tracking from RGB-D video sequences is a challenging problem due to the combination of changing viewpoints, motion, and occlusions over time. We observe that having the complete geometry of objects aids in their tracking, and thus propose to jointly infer the complete geometry of objects as well as track them, for rigidly moving objects over time. Our key insight is that inferring the complete geometry of the objects significantly helps in tracking. By hallucinating unseen regions of objects, we can obtain additional correspondences between the same instance, thus providing robust tracking even under strong change of appearance. From a sequence of RGB-D frames, we detect objects in each frame and learn to predict their complete object geometry as well as a dense correspondence mapping into a canonical space. This allows us to derive 6DoF poses for the objects in each frame, along with their correspondence between frames, providing robust object tracking across the RGB-D sequence. Experiments on both synthetic and real-world RGB-D data demonstrate that we achieve state-of-the-art performance on 3D multi-object tracking. Furthermore, we show that our object completion significantly helps tracking, providing an improvement of 8% in mean MOTA.
****
Multi-view Depth Estimation using Epipolar Spatio-Temporal Networks
Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8258-8267
We present a novel method for multi-view depth estimation from a single video, which is a critical task in various applications, such as perception, reconstruction and robot navigation. Although previous learning-based methods have demonstrated compelling results, most works estimate depth maps of individual video frames independently, without taking into consideration the strong geometric and temporal coherence among the frames. Moreover, current state-of-the-art (SOTA) models mostly adopt a fully 3D convolution network for cost regularization and therefore require high computational cost, thus limiting their deployment in real-world applications. Our method achieves temporally coherent depth estimation results by using a novel Epipolar Spatio-Temporal (EST) transformer to explicitly associate geometric and temporal correlation with multiple estimated depth maps. Furthermore, to reduce the computational cost, inspired by recent Mixture-of-Experts models, we design a compact hybrid network consisting of a 2D context-aware network and a 3D matching network which learn 2D context information and 3D disparity cues separately. Extensive experiments demonstrate that our method achieves higher accuracy in depth estimation and significant speedup than the SOTA methods.
****
AutoFlow: Learning a Better Training Set for Optical Flow
Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T. Freeman, Ce Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10093-10102
Synthetic datasets play a critical role in pre-training CNN models for optical flow, but they are painstaking to generate and hard to adapt to new applications. To automate the process, we present AutoFlow, a simple and effective method to render training data for optical flow that optimizes the performance of a model on a target dataset. AutoFlow takes a layered approach to render synthetic data,

where the motion, shape, and appearance of each layer are controlled by learnable hyperparameters. Experimental results show that AutoFlow achieves state-of-the-art accuracy in pre-training both PWC-Net and RAFT. Our code and data are available at autoflow-google.github.io.
****
LPSNet: A Lightweight Solution for Fast Panoptic Segmentation
Weixiang Hong, Qingpei Guo, Wei Zhang, Jingdong Chen, Wei Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16746-16754
Panoptic segmentation is a challenging task aiming to simultaneously segment objects (things) at instance level and background contents (stuff) at semantic level. Existing methods mostly utilize two-stage detection network to attain instance segmentation results, and fully convolutional network to produce semantic segmentation prediction. Post-processing or additional modules are required to handle the conflicts between the outputs from these two nets, which makes such methods suffer from low efficiency, heavy memory consumption and complicated implementation. To simplify the pipeline and decrease computation/memory cost, we propose an one-stage approach called Lightweight Panoptic Segmentation Network (LPSNet), which does not involve proposal, anchor or mask head. Instead, we predict bounding box and semantic category at each pixel upon the feature map produced by an augmented feature pyramid, and design a parameter-free head to merge the per-pixel bounding box and semantic prediction into panoptic segmentation output. Our LPSNet is not only efficient in computation and memory, but also accurate in panoptic segmentation. Comprehensive experiments on COCO, Cityscapes and Mapillary Vistas datasets demonstrate the promising effectiveness and efficiency of the proposed LPSNet.
****
You See What I Want You To See: Exploring Targeted Black-Box Transferability Attack for Hash-Based Image Retrieval Systems
Yanru Xiao, Cong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1934-1943
With the large multimedia content online, deep hashing has become a popular method for efficient image retrieval and storage. However, by inheriting the algorithmic backend from softmax classification, these techniques are vulnerable to the well-known adversarial examples as well. The massive collection of online images into the database also opens up new attack vectors. Attackers can embed adversarial images into the database and target specific categories to be retrieved by user queries. In this paper, we start from an adversarial standpoint to explore and enhance the capacity of targeted black-box transferability attack for deep hashing. We motivate this work by a series of empirical studies to see the unique challenges in image retrieval. We study the relations between adversarial subspace and black-box transferability via utilizing random noise as a proxy. Then we develop a new attack that is simultaneously adversarial and robust to noise to enhance transferability. Our experimental results demonstrate about 1.2-3x improvements of black-box transferability compared with the state-of-the-art mechanisms.
****
The Blessings of Unlabeled Background in Untrimmed Videos
Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6176-6185
Weakly-supervised Temporal Action Localization (WTAL) aims to detect the action segments with only video-level action labels in training. The key challenge is how to distinguish the action of interest segments from the background, which is unlabelled even on the video-level. While previous works treat the background as "curses", we consider it as "blessings". Specifically, we first use causal analysis to point out that the common localization errors are due to the unobserved confounder that resides ubiquitously in visual recognition. Then, we propose a Temporal Smoothing PCA-based (TS-PCA) deconfounder, which exploits the unlabelled background to model an observed substitute for the unobserved confounder, to re

move the confounding effect. Note that the proposed deconfounder is model-agnostic and non-intrusive, and hence can be applied in any WTAL method without model re-designs. Through extensive experiments on four state-of-the-art WTAL methods, we show that the deconfounder can improve all of them on the public datasets: THUMOS-14 and ActivityNet-1.3.
****
Autoregressive Stylized Motion Synthesis With Generative Flow
Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, Yong-Jin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13612-13621
Motion style transfer is an important problem in many computer graphics and computer vision applications, including human animation, games, and robotics. Most existing deep learning methods for this problem are supervised and trained by registered motion pairs. In addition, these methods are often limited to yielding a deterministic output, given a pair of style and content motions. In this paper, we propose an unsupervised approach for motion style transfer by synthesizing stylized motions autoregressively using a generative flow model M. M is trained to maximize the exact likelihood of a collection of unlabeled motions, based on an autoregressive context of poses in previous frames and a control signal representing the movement of a root joint. Thanks to invertible flow transformations, latent codes that encode deep properties of motion styles are efficiently inferred by M. By combining the latent codes (from an input style motion S) with the autoregressive context and control signal (from an input content motion C), M outputs a stylized motion which transfers style from S to C. Moreover, our model is probabilistic and is able to generate various plausible motions with a specific style. We evaluate the proposed model on motion capture datasets containing different human motion styles. Experiment results show that our model outperforms the state-of-the-art methods, despite not requiring manually labeled training data.
****
Improving Multiple Object Tracking With Single Object Tracking
Linyu Zheng, Ming Tang, Yingying Chen, Guibo Zhu, Jinqiao Wang, Hanqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2453-2462
Despite considerable similarities between multiple object tracking (MOT) and single object tracking (SOT) tasks, modern MOT methods have not benefited from the development of SOT ones to achieve satisfactory performance. The major reason for this situation is that it is inappropriate and inefficient to apply multiple SOT models directly to the MOT task, although advanced SOT methods are of the strong discriminative power and can run at fast speeds. In this paper, we propose a novel and end-to-end trainable MOT architecture that extends CenterNet by adding an SOT branch for tracking objects in parallel with the existing branch for object detection, allowing the MOT task to benefit from the strong discriminative power of SOT methods in an effective and efficient way. Unlike most existing SOT methods which learn to distinguish the target object from its local backgrounds, the added SOT branch trains a separate SOT model per target online to distinguish the target from its surrounding targets, assigning SOT models the novel discrimination. Moreover, similar to the detection branch, the SOT branch treats objects as points, making its online learning efficient even if multiple targets are processed simultaneously. Without tricks, the proposed tracker achieves MOTAs of 0.710 and 0.686, IDF1s of 0.719 and 0.714, on MOT17 and MOT20 benchmarks, respectively, while running at 16 FPS on MOT17.
****
Memory Oriented Transfer Learning for Semi-Supervised Image Deraining
Huaibo Huang, Aijing Yu, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7732-7741
Deep learning based methods have shown dramatic improvements in image rain removal by using large-scale paired data of synthetic datasets. However, due to the various appearances of real rain streaks that may be different from those in the synthetic training data, it is challenging to directly extend existing methods t

o the real-world scenes. To address this issue, we propose a memory-oriented semi-supervised (MOSS) method which enables the network to explore and exploit the properties of rain streaks from both synthetic and real data. The key aspect of our method is designing an encoder-decoder neural network that is augmented with a self-supervised memory module, where items in the memory record the prototypical patterns of rain degradations and are updated in a self-supervised way. Consequently, the rainy styles can be comprehensively derived from synthetic or real-world degraded images without the need for clean labels. Furthermore, we present a self-training mechanism that attempts to transfer deraining knowledge from supervised rain removal to unsupervised cases. An additional target network, which is updated with an exponential moving average of the online deraining network, is utilized to produce pseudo-labels for unlabeled rainy images. Meanwhile, the deraining network is optimized with supervised objectives on both synthetic paired data and pseudo-paired noisy data. Extensive experiments show that the proposed method achieves more appealing results not only on limited labeled data but also on unlabeled real-world images than recent state-of-the-art methods.
****
Instance Localization for Self-Supervised Detection Pretraining
Ceyuan Yang, Zhirong Wu, Bolei Zhou, Stephen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3987-3996
Prior research on self-supervised learning has led to considerable progress on image classification, but often with degraded transfer performance on object detection. The objective of this paper is to advance self-supervised pretrained models specifically for object detection. Based on the inherent difference between classification and detection, we propose a new self-supervised pretext task, called instance localization. Image instances are pasted at various locations and scales onto background images. The pretext task is to predict the instance category given the composited images as well as the foreground bounding boxes. We show that integration of bounding boxes into pretraining promotes better alignment between convolutional features and region boxes. In addition, we propose an augmentation method on the bounding boxes to further enhance this feature alignment. As a result, our model becomes weaker at Imagenet semantic classification but stronger at image patch localization, with an overall stronger pretrained model for object detection. Experimental results demonstrate that our approach yields state-of-the-art transfer learning results for object detection on PASCAL VOC and MSCOCO.
****
Adaptive Methods for Real-World Domain Generalization
Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, Dhruv Mahajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14340-14349
Invariant approaches have been remarkably successful in tackling the problem of domain generalization, where the objective is to perform inference on data distributions different from those used in training. In our work, we investigate whether it is possible to leverage domain information from the unseen test samples themselves. We propose a domain-adaptive approach consisting of two steps: a) we first learn a discriminative domain embedding from unsupervised training examples, and b) use this domain embedding as supplementary information to build a domain-adaptive model, that takes both the input as well as its domain into account while making predictions. For unseen domains, our method simply uses few unlabelled test examples to construct the domain embedding. This enables adaptive classification on any unseen domain. Our approach achieves state-of-the-art performance on various domain generalization benchmarks. In addition, we introduce the first real-world, large-scale domain generalization benchmark, Geo-YFCC, containing 1.1M samples over 40 training, 7 validation and 15 test domains, orders of magnitude larger than prior work. We show that the existing approaches either do not scale to this dataset or underperform compared to the simple baseline of training a model on the union of data from all training domains. In contrast, our approach achieves a significant 1% improvement.
****

Deep Animation Video Interpolation in the Wild
Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6587-6595
In the animation industry, cartoon videos are usually produced at low frame rate since hand drawing of such frames is costly and time-consuming. Therefore, it is desirable to develop computational models that can automatically interpolate the in-between animation frames. However, existing video interpolation methods fail to produce satisfying results on animation data. Compared to natural videos, animation videos possess two unique characteristics that make frame interpolation difficult: 1) cartoons comprise lines and smooth color pieces. The smooth areas lack textures and make it difficult to estimate accurate motions on animation videos. 2) cartoons express stories via exaggeration. Some of the motions are non-linear and extremely large. In this work, we formally define and study the animation video interpolation problem for the first time. To address the aforementioned challenges, we propose an effective framework, AnimeInterp, with two dedicated modules in a coarse-to-fine manner. Specifically, 1) Segment-Guided Matching resolves the "lack of textures" challenge by exploiting global matching among color pieces that are piece-wise coherent. 2) Recurrent Flow Refinement resolves the "non-linear and extremely large motion" challenge by recurrent predictions using a transformer-like architecture. To facilitate comprehensive training and evaluations, we build a large-scale animation triplet dataset, ATD-12K, which comprises 12,000 triplets with rich annotations. Extensive experiments demonstrate that our approach outperforms existing state-of-the-art interpolation methods for animation videos. Notably, AnimeInterp shows favorable perceptual quality and robustness for animation scenarios in the wild. The proposed dataset and code are available at https://github.com/lisiyao21/AnimeInterp/.
****
Isometric Multi-Shape Matching
Maolin Gao, Zorah Lahner, Johan Thunberg, Daniel Cremers, Florian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14183-14193
Finding correspondences between shapes is a fundamental problem in computer vision and graphics, which is relevant for many applications, including 3D reconstruction, object tracking, and style transfer. The vast majority of correspondence methods aim to find a solution between pairs of shapes, even if multiple instances of the same class are available. While isometries are often studied in shape correspondence problems, they have not been considered explicitly in the multi-matching setting. This paper closes this gap by proposing a novel optimisation formulation for isometric multi-shape matching. We present a suitable optimisation algorithm for solving our formulation and provide a convergence and complexity analysis. Moreover, our algorithm obtains multi-matchings that are cycle-consistent without having to explicitly enforce cycle-consistency constraints. We demonstrate the superior performance of our method on various datasets and set the new state-of-the-art in isometric multi-shape matching.
****
Spatially Consistent Representation Learning
Byungseok Roh, Wuhyun Shin, Ildoo Kim, Sungwoong Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1144-1153
Self-supervised learning has been widely used to obtain transferrable representations from unlabeled images. Especially, recent contrastive learning methods have shown impressive performances on downstream image classification tasks. While these contrastive methods mainly focus on generating invariant global representations at the image-level under semantic-preserving transformations, they are prone to overlook spatial consistency of local representations and therefore have a limitation in pretraining for localization tasks such as object detection and instance segmentation. Moreover, aggressively cropped views used in existing contrastive methods can minimize representation distances between the semantically different regions of a single image. In this paper, we propose a spatially consis

tent representation learning algorithm (SCRL) for multi-object and location-specific tasks. In particular, we devise a novel self-supervised objective that tries to produce coherent spatial representations of a randomly cropped local region according to geometric translations and zooming operations. On various downstream localization tasks with benchmark datasets, the proposed SCRL shows significant performance improvements over the image-level supervised pretraining as well as the state-of-the-art self-supervised learning methods.
****
Semantic Scene Completion via Integrating Instances and Scene In-the-Loop
Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 324-333
Semantic Scene Completion aims at reconstructing a complete 3D scene with precise voxel-wise semantics from a single-view depth or RGBD image. It is a crucial but challenging problem for indoor scene understanding. In this work, we present a novel framework named Scene-Instance-Scene Network (SISNet), which takes advantages of both instance and scene level semantic information. Our method is capable of inferring fine-grained shape details as well as nearby objects whose semantic categories are easily mixedup. The key insight is that we decouple the instances from a coarsely completed semantic scene instead of a raw input image to guide the reconstruction of instances and the overall scene. SISNet conducts iterative scene-to-instance (SI) and instance-to-scene (IS) semantic completion. Specifically, the SI is able to encode objects' surrounding context for effectively decoupling instances from the scene and each instance could be voxelized into higher resolution to capture finer details. With IS, fine-grained instance information can be integrated back into the 3D scene and thus leads to more accurate semantic scene completion. Utilizing such an iterative mechanism, the scene and instance completion benefits each other to achieve higher completion accuracy. Extensively experiments show that our proposed method consistently outperforms state-of-the-art methods on both real NYU, NYUCAD and synthetic SUNCG-RGBD datasets. The code and the supplementary material will be available at https://github.com/yjcaimeow/SISNet.
****
Efficient Deformable Shape Correspondence via Multiscale Spectral Manifold Wavelets Preservation
Ling Hu, Qinsong Li, Shengjun Liu, Xinru Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14536-14545
The functional map framework has proven to be extremely effective for representing dense correspondences between deformable shapes. A key step in this framework is to formulate suitable preservation constraints to encode the geometric information that must be preserved by the unknown map. For this issue, we construct novel and powerful constraints to determine the functional map, where multiscale spectral manifold wavelets are required to be preserved at each scale correspondingly. Such constraints allow us to extract significantly more information than previous methods, especially those based on descriptor preservation constraints, and strongly ensure the isometric property of the map. In addition, we also propose a remarkable efficient iterative method to alternatively update the functional maps and pointwise maps. Moreover, when we use the tight wavelet frames in iterations, the computation of the functional maps boils down to a simple filtering procedure with low-pass and various band-pass filters, which avoids time-consuming solving large systems of linear equations commonly presented in functional maps. We demonstrate on a wide variety of experiments with different datasets that our approach achieves significant improvements both in the shape correspondence quality and the computing efficiency.
****
TearingNet: Point Cloud Autoencoder To Learn Topology-Friendly Representations
Jiahao Pang, Duanshun Li, Dong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7453-7462
Topology matters. Despite the recent success of point cloud processing with geometric deep learning, it remains arduous to capture the complex topologies of poi

nt cloud data with a learning model. Given a point cloud dataset containing objects with various genera, or scenes with multiple objects, we propose an autoencoder, TearingNet, which tackles the challenging task of representing the point clouds using a fixed-length descriptor. Unlike existing works directly deforming predefined primitives of genus zero (e.g., a 2D square patch) to an object-level point cloud, our TearingNet is characterized by a proposed Tearing network module and a Folding network module interacting with each other iteratively. Particularly, the Tearing network module learns the point cloud topology explicitly. By breaking the edges of a primitive graph, it tears the graph into patches or with holes to emulate the topology of a target point cloud, leading to faithful reconstructions. Experimentation shows the superiority of our proposal in terms of reconstructing point clouds as well as generating more topology-friendly representations than benchmarks.
****
Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics
Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, Wenqi Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16469-16477
Neural network ensembles are gaining popularity by harnessing the complementary wisdom of multiple base models. Ensemble teams with high diversity promote high failure independence, which is effective for boosting the overall ensemble accuracy. This paper provides an in-depth study on how to design and compute ensemble diversity, which can capture the complementary decision capacity of ensemble member models. We make three original contributions. First, we revisit the ensemble diversity metrics in the literature and analyze the inherent problems of poor correlation between ensemble diversity and ensemble accuracy, which leads to the low quality ensemble selection using such diversity metrics. Second, instead of computing diversity scores for ensemble teams of different sizes using the same criteria, we introduce focal model based ensemble diversity metrics, coined as FQ-diversity metrics. Our new metrics significantly improve the intrinsic correlation between high ensemble diversity and high ensemble accuracy. Third, we introduce a diversity fusion method, coined as the EQ-diversity metric, by integrating the top three most representative FQ-diversity metrics. Comprehensive experiments on two benchmark datasets (CIFAR-10 and ImageNet) show that our FQ and EQ diversity metrics are effective for selecting high diversity ensemble teams to boost overall ensemble accuracy.
****
WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition
Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10492-10502
In this paper, we contribute a new million-scale face benchmark containing noisy 4M identities/260M faces (WebFace260M) and cleaned 2M identities/42M faces (WebFace42M) training data, as well as an elaborately designed time-constrained evaluation protocol. Firstly, we collect 4M name list and download 260M faces from the Internet. Then, a Cleaning Automatically utilizing Self-Training (CAST) pipeline is devised to purify the tremendous WebFace260M, which is efficient and scalable. To the best of our knowledge, the cleaned WebFace42M is the largest public face recognition training set and we expect to close the data gap between academia and industry. Referring to practical scenarios, Face Recognition Under Inference Time conStraint (FRUITS) protocol and a test set are constructed to comprehensively evaluate face matchers. Equipped with this benchmark, we delve into million-scale face recognition problems. A distributed framework is developed to train face recognition models efficiently without tampering with the performance. Empowered by WebFace42M, we reduce relative 40% failure rate on the challenging IJB-C set, and rank the 3rd among 430 entries on NIST-FRVT. Even 10% data (WebFace4M) shows superior performance compared with public training set. Furthermore, comprehensive baselines are established on our rich-attribute test set under FRUITS-100ms/500ms/1000ms protocol, including MobileNet, EfficientNet, AttentionNe

t, ResNet, SENet, ResNeXt and RegNet families. Benchmark website is https://www.
face-benchmark.org.
****
RSN: Range Sparse Net for Efficient, Accurate LiDAR 3D Object Detection
Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang,
Cristian Sminchisescu, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5725-5734
The detection of 3D objects from LiDAR data is a critical component in most auto
nomous driving systems. Safe, high speed driving needs larger detection ranges,
which are enabled by new LiDARs. These larger detection ranges require more effi
cient and accurate detection models. Towards this goal, we propose Range Sparse
Net (RSN) - a simple, efficient, and accurate 3D object detector - in order to t
ackle real time 3D object detection in this extended detection regime. RSN predi
cts foreground points from range images and applies sparse convolutions on the s
elected fore-ground points to detect objects. The lightweight 2D convolutions on
 dense range images results in significantly fewer selected foreground points, t
hus enabling the later sparse convolutions in RSN to efficiently operate. Combin
ing features from the range image further enhance detection ac-curacy. RSN runs
at more than 60 frames per second on a 150mx150m detection region on Waymo Open
Dataset (WOD) while being more accurate than previously published detectors. RSN
 is ranked first in the WOD leaderboard based on the APH/LEVEL1 metrics for LiDA
R-based pedestrian and vehicle detection, while being several times faster than
alternatives.
****
Labeled From Unlabeled: Exploiting Unlabeled Data for Few-Shot Deep HDR Deghosti
ng
K. Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R. Venkatesh Babu, Rama Krish
na Sai S Gorthi; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2021, pp. 4875-4885
High Dynamic Range (HDR) deghosting is an indispensable tool in capturing wide d
ynamic range scenes without ghosting artifacts. Recently, convolutional neural n
etworks (CNNs) have shown tremendous success in HDR deghosting. However, CNN-bas
ed HDR deghosting methods require collecting large datasets with ground truth, w
hich is a tedious and time-consuming process. This paper proposes a pioneering w
ork by introducing zero and few-shot learning strategies for data-efficient HDR
deghosting. Our approach consists of two stages of training. In stage one, we tr
ain the model with few labeled (5 or less) dynamic samples and a pool of unlabel
ed samples with a self-supervised loss. We use the trained model to predict HDRs
 for the unlabeled samples. To derive data for the next stage of training, we pr
opose a novel method for generating corresponding dynamic inputs from the predic
ted HDRs of unlabeled data. The generated artificial dynamic inputs and predicte
d HDRs are used as paired labeled data. In stage two, we finetune the model with
 the original few labeled data and artificially generated labeled data. Our few-
shot approach outperforms many fully-supervised methods in two publicly availabl
e datasets, using as little as five labeled dynamic samples.
****
Convolutional Dynamic Alignment Networks for Interpretable Classifications
Moritz Bohle, Mario Fritz, Bernt Schiele; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10029-10038
We introduce a new family of neural network models called Convolutional Dynamic
Alignment Networks (CoDA-Nets), which are performant classifiers with a high deg
ree of inherent interpretability. Their core building blocks are Dynamic Alignme
nt Units (DAUs), which linearly transform their input with weight vectors that d
ynamically align with task-relevant patterns. As a result, CoDA-Nets model the c
lassification prediction through a series of input-dependent linear transformati
ons, allowing for linear decomposition of the output into individual input contr
ibutions. Given the alignment of the DAUs, the resulting contribution maps align
 with discriminative input patterns. These model-inherent decompositions are of
high visual quality and outperform existing attribution methods under quantitati
ve metrics. Further, CoDA-Nets constitute performant classifiers, achieving on p

ar results to ResNet and VGG models on e.g. CIFAR-10 and TinyImagenet.
****
EDNet: Efficient Disparity Estimation With Cost Volume Combination and Attention-Based Spatial Residual
Songyan Zhang, Zhicheng Wang, Qiang Wang, Jinshuo Zhang, Gang Wei, Xiaowen Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5433-5442
Existing state-of-the-art disparity estimation works mostly leverage the 4D concatenation volume and construct a very deep 3D convolution neural network (CNN) for disparity regression, which is inefficient due to the high memory consumption and slow inference speed. In this paper, we propose a network named EDNet for efficient disparity estimation. Firstly, we construct a combined volume which incorporates contextual information from the squeezed concatenation volume and feature similarity measurement from the correlation volume. The combined volume can be next aggregated by 2D convolutions which are faster and require less memory than 3D convolutions. Secondly, we propose an attention-based spatial residual module to generate attention-aware residual features. The attention mechanism is applied to provide intuitive spatial evidence about inaccurate regions with the help of error maps at multiple scales and thus improve the residual learning efficiency. Extensive experiments on the Scene Flow and KITTI datasets show that EDNet outperforms the previous 3D CNN based works and achieves state-of-the-art performance with significantly faster speed and less memory consumption.
****
Unsupervised Visual Representation Learning by Tracking Patches in Video
Guangting Wang, Yizhou Zhou, Chong Luo, Wenxuan Xie, Wenjun Zeng, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2563-2572
Inspired by the fact that human eyes continue to develop tracking ability in early and middle childhood, we propose to use tracking as a proxy task for a computer vision system to learn the visual representations. Modelled on the Catch game played by the children, we design a Catch-the-Patch (CtP) game for a 3D-CNN model to learn visual representations that would help with video-related tasks. In the proposed pretraining framework, we cut an image patch from a given video and let it scale and move according to a pre-set trajectory. The proxy task is to estimate the position and size of the image patch in a sequence of video frames, given only the target bounding box in the first frame. We discover that using multiple image patches simultaneously brings clear benefits. We further increase the difficulty of the game by randomly making patches invisible. Extensive experiments on mainstream benchmarks demonstrate the superior performance of CtP against other video pretraining methods. In addition, CtP-pretrained features are less sensitive to domain gaps than those trained by a supervised action recognition task. When both trained on Kinetics-400, we are pleasantly surprised to find that CtP-pretrained representation achieves much higher action classification accuracy than its fully supervised counterpart on Something-Something dataset.
****
Wasserstein Contrastive Representation Distillation
Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, Lawrence Carin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16296-16305
The primary goal of knowledge distillation (KD) is to encapsulate the information of a model learned from a teacher network into a student network, with the latter being more compact than the former. Existing work, e.g., using Kullback-Leibler divergence for distillation, may fail to capture important structural knowledge in the teacher network and often lacks the ability for feature generalization, particularly in situations when teacher and student are built to address different classification tasks. We propose Wasserstein Contrastive Representation Distillation (WCoRD), which leverages both primal and dual forms of Wasserstein distance for KD. The dual form is used for global knowledge transfer, yielding a contrastive learning objective that maximizes the lower bound of mutual information between the teacher and the student networks. The primal form is used for loc

al contrastive knowledge transfer within a mini-batch, effectively matching the distributions of features between the teacher and the student networks. Experiments demonstrate that the proposed WCoRD method outperforms state-of-the-art approaches on privileged information distillation, model compression and cross-modal transfer.
****

Learnable Companding Quantization for Accurate Low-Bit Neural Networks
Kohei Yamamoto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5029-5038
Quantizing deep neural networks is an effective method for reducing memory consumption and improving inference speed, and is thus useful for implementation in resource-constrained devices. However, it is still hard for extremely low-bit models to achieve accuracy comparable with that of full-precision models. To address this issue, we propose learnable companding quantization (LCQ) as a novel non-uniform quantization method for 2-, 3-, and 4-bit models. LCQ jointly optimizes model weights and learnable companding functions that can flexibly and non-uniformly control the quantization levels of weights and activations. We also present a new weight normalization technique that allows more stable training for quantization. Experimental results show that LCQ outperforms conventional state-of-the-art methods and narrows the gap between quantized and full-precision models for image classification and object detection tasks. Notably, the 2-bit ResNet-50 model on ImageNet achieves top-1 accuracy of 75.1% and reduces the gap to 1.7%, allowing LCQ to further exploit the potential of non-uniform quantization.
****

FaceInpainter: High Fidelity Face Adaptation to Heterogeneous Domains
Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, Ran He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5089-5098
In this work, we propose a novel two-stage framework named FaceInpainter to implement controllable Identity-Guided Face Inpainting (IGFI) under heterogeneous domains. Concretely, by explicitly disentangling foreground and background of the target face, the first stage focuses on adaptive face fitting to the fixed background via a Styled Face Inpainting Network (SFI-Net), with 3D priors and texture code of the target, as well as identity factor of the source face. It is challenging to deal with the inconsistency between the new identity of the source and the original background of the target, concerning the face shape and appearance on the fused boundary. The second stage consists of a Joint Refinement Network (JR-Net) to refine the swapped face. It leverages AdaIN considering identity and multi-scale texture codes, for feature transformation of the decoded face from SFI-Net with facial occlusions. We adopt the contextual loss to implicitly preserve the attributes, encouraging face deformation and fewer texture distortions. Experimental results demonstrate that our approach handles high-quality identity adaptation to heterogeneous domains, exhibiting the competitive performance compared with state-of-the-art methods concerning both attribute and identity fidelity.
****

How Robust Are Randomized Smoothing Based Defenses to Data Poisoning?
Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, Jihun Hamm; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13244-13253
Predictions of certifiably robust classifiers remain constant in a neighborhood of a point, making them resilient to test-time attacks with a guarantee. In this work, we present a previously unrecognized threat to robust machine learning models that highlights the importance of training-data quality in achieving high certified adversarial robustness. Specifically, we propose a novel bilevel optimization-based data poisoning attack that degrades the robustness guarantees of certifiably robust classifiers. Unlike other poisoning attacks that reduce the accuracy of the poisoned models on a small set of target points, our attack reduces the average certified radius (ACR) of an entire target class in the dataset. Moreover, our attack is effective even when the victim trains the models from scra

tch using state-of-the-art robust training methods such as Gaussian data augment
ation [??], MACER [??], and SmoothAdv [??] that achieve high certified adversari
al robustness. To make the attack harder to detect, we use clean-label poisoning
 points with imperceptible distortions. The effectiveness of the proposed method
 is evaluated by poisoning MNIST and CIFAR10 datasets and training deep neural n
etworks using previously mentioned training methods and certifying the robustnes
s with randomized smoothing. The ACR of the target class, for models trained on
generated poison data, can be reduced by more than 30%. Moreover, the poisoned d
ata is transferable to models trained with different training methods and models
 with different architectures.
****
Deep Learning in Latent Space for Video Prediction and Compression
Bowen Liu, Yu Chen, Shiyu Liu, Hun-Seok Kim; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 701-710
Learning-based video compression has achieved substantial progress during recent
 years. The most influential approaches adopt deep neural networks (DNNs) to rem
ove spatial and temporal redundancies by finding the appropriate lower-dimension
al representations of frames in the video. We propose a novel DNN based framewor
k that predicts and compresses video sequences in the latent vector space. The p
roposed method first learns the efficient lower-dimensional latent space represe
ntation of each video frame and then performs inter-frame prediction in that lat
ent domain. The proposed latent domain compression of individual frames is obtai
ned by a deep autoencoder trained with a generative adversarial network (GAN). T
o exploit the temporal correlation within the video frame sequence, we employ a
convolutional long short-term memory (ConvLSTM) network to predict the latent ve
ctor representation of the future frame. We demonstrate our method with two appl
ications; video compression and abnormal event detection that share the identica
l latent frame prediction network. The proposed method exhibits superior or comp
etitive performance compared to the state-of-the-art algorithms specifically des
igned for either video compression or anomaly detection.
****
PWCLO-Net: Deep LiDAR Odometry in 3D Point Clouds Using Hierarchical Embedding M
ask Optimization
Guangming Wang, Xinrui Wu, Zhe Liu, Hesheng Wang; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15910-1591
9
A novel 3D point cloud learning model for deep LiDAR odometry, named PWCLO-Net,
using hierarchical embedding mask optimization is proposed in this paper. In thi
s model, the Pyramid, Warping, and Cost volume (PWC) structure for the LiDAR odo
metry task is built to refine the estimated pose in a coarse-to-fine approach hi
erarchically. An attentive cost volume is built to associate two point clouds an
d obtain embedding motion patterns. Then, a novel trainable embedding mask is pr
oposed to weigh the local motion patterns of all points to regress the overall p
ose and filter outlier points. The estimated current pose is used to warp the fi
rst point cloud to bridge the distance to the second point cloud, and then the c
ost volume of the residual motion is built. At the same time, the embedding mask
 is optimized hierarchically from coarse to fine to obtain more accurate filteri
ng information for pose refinement. The trainable pose warp-refinement process i
s iteratively used to make the pose estimation more robust for outliers. The sup
erior performance and effectiveness of our LiDAR odometry model are demonstrated
 on KITTI odometry dataset. Our method outperforms all recent learning-based met
hods and outperforms the geometry-based approach, LOAM with mapping optimization
, on most sequences of KITTI odometry dataset. Our source codes will be released
 on https://github.com/IRMVLab/PWCLONet.
****
ORDisCo: Effective and Efficient Usage of Incremental Unlabeled Data for Semi-Su
pervised Continual Learning
Liyuan Wang, Kuo Yang, Chongxuan Li, Lanqing Hong, Zhenguo Li, Jun Zhu; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2021, pp. 5383-5392

Continual learning usually assumes the incoming data are fully labeled, which might not be applicable in real applications. In this work, we consider semi-supervised continual learning (SSCL) that incrementally learns from partially labeled data. Observing that existing continual learning methods lack the ability to continually exploit the unlabeled data, we propose deep Online Replay with Discriminator Consistency (ORDisCo) to interdependently learn a classifier with a conditional generative adversarial network (GAN), which continually passes the learned data distribution to the classifier. In particular, ORDisCo replays data sampled from the conditional generator to the classifier in an online manner, exploiting unlabeled data in a time- and storage-efficient way. Further, to explicitly overcome the catastrophic forgetting of unlabeled data, we selectively stabilize parameters of the discriminator that are important for discriminating the pairs of old unlabeled data and their pseudo-labels predicted by the classifier. We extensively evaluate ORDisCo on various semi-supervised learning benchmark datasets for SSCL, and show that ORDisCo achieves significant performance improvement on SVHN, CIFAR10 and Tiny-ImageNet, compared to strong baselines.
****

Dynamic Region-Aware Convolution

Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8064-8073

We propose a new convolution called Dynamic Region-Aware Convolution (DRConv), which can automatically assign multiple filters to corresponding spatial regions where features have similar representation. In this way, DRConv outperforms standard convolution in modeling semantic variations. Standard convolutional layer can increase the number of filers to extract more visual elements but results in high computational cost. More gracefully, our DRConv transfers the increasing channel-wise filters to spatial dimension with learnable instructor, which not only improve representation ability of convolution, but also maintains computational cost and the translation-invariance as standard convolution dose. DRConv is an effective and elegant method for handling complex and variable spatial information distribution. It can substitute standard convolution in any existing networks for its plug-and-play property, especially to power convolution layers in efficient networks. We evaluate DRConv on a wide range of models (MobileNet series, ShuffleNetV2, etc.) and tasks (Classification, Face Recognition, Detection and Segmentation). On ImageNet classification, DRConv-based ShuffleNetV2-0.5x achieves state-of-the-art performance of 67.1% at 46M multiply-adds level with 6.3% relative improvement.
****

Explore Image Deblurring via Encoded Blur Kernel Space

Phong Tran, Anh Tuan Tran, Quynh Phung, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11956-11965

This paper introduces a method to encode the blur operators of an arbitrary dataset of sharp-blur image pairs into a blur kernel space. Assuming the encoded kernel space is close enough to in-the-wild blur operators, we propose an alternating optimization algorithm for blind image deblurring. It approximates an unseen blur operator by a kernel in the encoded space and searches for the corresponding sharp image. Unlike recent deep-learning-based methods, our system can handle unseen blur kernel, while avoiding using complicated handcrafted priors on the blur operator often found in classical methods. Due to the method's design, the encoded kernel space is fully differentiable, thus can be easily adopted in deep neural network models. Moreover, our method can be used for blur synthesis by transferring existing blur operators from a given dataset into a new domain. Finally, we provide experimental results to confirm the effectiveness of the proposed method.
****

BCNet: Searching for Network Width With Bilaterally Coupled Network

Xiu Su, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021

, pp. 2175-2184
Searching for a more compact network width recently serves as an effective way o
f channel pruning for the deployment of convolutional neural networks (CNNs) und
er hardware constraints. To fulfill the searching, a one-shot supernet is usuall
y leveraged to efficiently evaluate the performance \wrt different network width
s. However, current methods mainly follow a unilaterally augmented (UA) principl
e for the evaluation of each width, which induces the training unfairness of cha
nnels in supernet. In this paper, we introduce a new supernet called Bilaterally
 Coupled Network (BCNet) to address this issue. In BCNet, each channel is fairly
 trained and responsible for the same amount of network widths, thus each networ
k width can be evaluated more accurately. Besides, we leverage a stochastic comp
lementary strategy for training the BCNet, and propose a prior initial populatio
n sampling method to boost the performance of the evolutionary search. Extensive
 experiments on benchmark CIFAR-10 and ImageNet datasets indicate that our metho
d can achieve state-of-the-art or competing performance over other baseline meth
ods. Moreover, our method turns out to further boost the performance of NAS mode
ls by refining their network widths. For example, with the same FLOPs budget, ou
r obtained EfficientNet-B0 achieves 77.36% Top-1 accuracy on ImageNet dataset, s
urpassing the performance of original setting by 0.48%.
****
Camera Pose Matters: Improving Depth Prediction by Mitigating Pose Distribution
Bias
Yunhan Zhao, Shu Kong, Charless Fowlkes; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15759-15768
Monocular depth predictors are typically trained on large-scale training sets wh
ich are naturally biased w.r.t the distribution of camera poses. As a result, tr
ained predictors fail to make reliable depth predictions for testing examples ca
ptured under uncommon camera poses. To address this issue, we propose two novel
techniques that exploit the camera pose during training and prediction. First, w
e introduce a simple perspective-aware data augmentation that synthesizes new tr
aining examples with more diverse views by perturbing the existing ones in a geo
metrically consistent manner. Second, we propose a conditional model that exploi
ts the per-image camera pose as prior knowledge by encoding it as a part of the
input. We show that jointly applying the two methods improves depth prediction o
n images captured under uncommon and even never-before-seen camera poses. We sho
w that our methods improve performance when applied to a range of different pred
ictor architectures. Lastly, we show that explicitly encoding the camera pose di
stribution improves the generalization performance of a synthetically trained de
pth predictor when evaluated on real images.
****
Lipstick Ain't Enough: Beyond Color Matching for In-the-Wild Makeup Transfer
Thao Nguyen, Anh Tuan Tran, Minh Hoai; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13305-13314
Makeup transfer is the task of applying on a source face the makeup style from a
 reference image. Real-life makeups are diverse and wild, which cover not only c
olor-changing but also patterns, such as stickers, blushes, and jewelries. Howev
er, existing works overlooked the latter components and confined makeup transfer
 to color manipulation, focusing only on light makeup styles. In this work, we p
ropose a holistic makeup transfer framework that can handle all the mentioned ma
keup components. It consists of an improved color transfer branch and a novel pa
ttern transfer branch to learn all makeup properties, including color, shape, te
xture, and location. To train and evaluate such a system, we also introduce new
makeup datasets for real and synthetic extreme makeup. Experimental results show
 that our framework achieves the state of the art performance on both light and
extreme makeup styles.
****
Generative Interventions for Causal Learning
Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, Carl Vondrick;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2021, pp. 3947-3956

We introduce a framework for learning robust visual representations that generalize to new viewpoints, backgrounds, and scene contexts. Discriminative models often learn naturally occurring spurious correlations, which cause them to fail on images outside of the training distribution. In this paper, we show that we can steer generative models to manufacture interventions on features caused by confounding factors. Experiments, visualizations, and theoretical results show this method learns robust representations more consistent with the underlying causal relationships. Our approach improves performance on multiple datasets demanding out-of-distribution generalization, and we demonstrate state-of-the-art performance generalizing from ImageNet to ObjectNet dataset.
****

Graph Stacked Hourglass Networks for 3D Human Pose Estimation
Tianhan Xu, Wataru Takano; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16105-16114
In this paper, we propose a novel graph convolutional network architecture, Graph Stacked Hourglass Networks, for 2D-to-3D human pose estimation tasks. The proposed architecture consists of repeated encoder-decoder, in which graph-structured features are processed across three different scales of human skeletal representations. This multi-scale architecture enables the model to learn both local and global feature representations, which are critical for 3D human pose estimation. We also introduce a multi-level feature learning approach using different-depth intermediate features and show the performance improvements that result from exploiting multi-scale, multi-level feature representations. Extensive experiments are conducted to validate our approach, and the results show that our model outperforms the state-of-the-art.
****

Adaptive Aggregation Networks for Class-Incremental Learning
Yaoyao Liu, Bernt Schiele, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2544-2553
Class-Incremental Learning (CIL) aims to learn a classification model with the number of classes increasing phase-by-phase. An inherent problem in CIL is the stability-plasticity dilemma between the learning of old and new classes, i.e., high-plasticity models easily forget old classes, but high-stability models are weak to learn new classes. We alleviate this issue by proposing a novel network architecture called Adaptive Aggregation Networks (AANets), in which we explicitly build two types of residual blocks at each residual level (taking ResNet as the baseline architecture): a stable block and a plastic block. We aggregate the output feature maps from these two blocks and then feed the results to the next-level blocks. We adapt the aggregation weights in order to balance these two types of blocks, i.e., to balance stability and plasticity, dynamically. We conduct extensive experiments on three CIL benchmarks: CIFAR-100, ImageNet-Subset, and ImageNet, and show that many existing CIL methods can be straightforwardly incorporated into the architecture of AANets to boost their performances.
****

VS-Net: Voting With Segmentation for Visual Localization
Zhaoyang Huang, Han Zhou, Yijin Li, Bangbang Yang, Yan Xu, Xiaowei Zhou, Hujun Bao, Guofeng Zhang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6101-6111
Visual localization is of great importance in robotics and computer vision. Recently, scene coordinate regression based methods have shown good performance in visual localization in small static scenes. However, it still estimates camera poses from many inferior scene coordinates. To address this problem, we propose a novel visual localization framework that establishes 2D-to-3D correspondences between the query image and the 3D map with a series of learnable scene-specific landmarks. In the landmark generation stage, the 3D surfaces of the target scene are over-segmented into mosaic patches whose centers are regarded as the scene-specific landmarks. To robustly and accurately recover the scene-specific landmarks, we propose the Voting with Segmentation Network (VS-Net) to segment the pixels into different landmark patches with a segmentation branch and estimate the landmark locations within each patch with a landmark location voting branch. Sinc

e the number of landmarks in a scene may reach up to 5000, training a segmentati
on network with such a large number of classes is both computation and memory co
stly for the commonly used cross-entropy loss. We propose a novel prototype-base
d triplet loss with hard negative mining, which is able to train semantic segmen
tation networks with a large number of labels efficiently. Our proposed VS-Net i
s extensively tested on multiple public benchmarks and can outperform state-of-t
he-art visual localization methods. Code and models are available at https://git
hub.com/zju3dv/VS-Net.
****
Learning To Identify Correct 2D-2D Line Correspondences on Sphere
Haoang Li, Kai Chen, Ji Zhao, Jiangliu Wang, Pyojin Kim, Zhe Liu, Yun-Hui Liu; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2021, pp. 11743-11752
Given a set of putative 2D-2D line correspondences, we aim to identify correct m
atches. Existing methods exploit the geometric constraints. They are only applic
able to structured scenes with orthogonality, parallelism and coplanarity. In co
ntrast, we propose the first approach suitable for both structured and unstructu
red scenes. Instead of geometric constraint, we leverage the spatial regularity
on sphere. Specifically, we propose to map line correspondences into vectors tan
gent to sphere. We use these vectors to encode both angular and positional varia
tions of image lines, which is more reliable and concise than directly using inc
linations, midpoints or endpoints of image lines. Neighboring vectors mapped fro
m correct matches exhibit a spatial regularity called local trend consistency, r
egardless of the type of scenes. To encode this regularity, we design a neural n
etwork and also propose a novel loss function that enforces the smoothness const
raint of vector field. In addition, we establish a large real-world dataset for
image line matching. Experiments showed that our approach outperforms state-of-t
he-art ones in terms of accuracy, efficiency and robustness, and also leads to h
igh generalization.
****
Domain-Independent Dominance of Adaptive Methods
Pedro Savarese, David McAllester, Sudarshan Babu, Michael Maire; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
 pp. 16286-16295
From a simplified analysis of adaptive methods, we derive AvaGrad, a new optimiz
er which outperforms SGD on vision tasks when its adaptability is properly tuned
. We observe that the power of our method is partially explained by a decoupling
 of learning rate and adaptability, greatly simplifying hyperparameter search. I
n light of this observation, we demonstrate that, against conventional wisdom, A
dam can also outperform SGD on vision tasks, as long as the coupling between its
 learning rate and adaptability is taken into account. In practice, AvaGrad matc
hes the best results, as measured by generalization accuracy, delivered by any e
xisting optimizer (SGD or adaptive) across image classification (CIFAR, ImageNet
) and character-level language modelling (Penn Treebank) tasks. When training GA
Ns, AvaGrad improves upon existing optimizers.
****
What if We Only Use Real Datasets for Scene Text Recognition? Toward Scene Text
Recognition With Fewer Labels
Jeonghun Baek, Yusuke Matsui, Kiyoharu Aizawa; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3113-3122
Scene text recognition (STR) task has a common practice: All state-of-the-art ST
R models are trained on large synthetic data. In contrast to this practice, trai
ning STR models only on fewer real labels (STR with fewer labels) is important w
hen we have to train STR models without synthetic data: for handwritten or artis
tic texts that are difficult to generate synthetically and for languages other t
han English for which we do not always have synthetic data. However, there has b
een implicit common knowledge that training STR models on real data is nearly im
possible because real data is insufficient. We consider that this common knowled
ge has obstructed the study of STR with fewer labels. In this work, we would lik
e to reactivate STR with fewer labels by disproving the common knowledge. We con

solidate recently accumulated public real data and show that we can train STR mo dels satisfactorily only with real labeled data. Subsequently, we find simple da ta augmentation to fully exploit real data. Furthermore, we improve the models b y collecting unlabeled data and introducing semi- and self-supervised methods. A s a result, we obtain a competitive model to state-of-the-art methods. To the be st of our knowledge, this is the first study that 1) shows sufficient performanc e by only using real labels and 2) introduces semi- and self-supervised methods into STR with fewer labels. Our code and data are available.
****
Incremental Learning via Rate Reduction
Ziyang Wu, Christina Baek, Chong You, Yi Ma; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1125-1133
Current deep learning architectures suffer from catastrophic forgetting, a failu re to retain knowledge of previously learned classes when incrementally trained on new classes. The fundamental roadblock faced by deep learning methods is that the models are optimized as "black boxes", making it difficult to properly adju st the model parameters to preserve knowledge about previously seen data. To ove rcome the problem of catastrophic forgetting, we propose utilizing an alternativ e "white box" architecture derived from the principle of rate reduction, where e ach layer of the network is explicitly computed without back propagation. Under this paradigm, we demonstrate that, given a pretrained network and new data clas ses, our approach can provably construct a new network that emulates joint train ing with all past and new classes. Finally, our experiments show that our propos ed learning algorithm observes significantly less decay in classification perfor mance, outperforming state of the art methods on MNIST and CIFAR-10 by a large m argin and justifying the use of "white box" algorithms for incremental learning even for sufficiently complex image data.
****
Neural Descent for Visual 3D Human Pose and Shape
Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul S ukthankar, Cristian Sminchisescu; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2021, pp. 14484-14493
We present deep neural network methodology to reconstruct the 3d pose and shape of people, including hand gestures and facial expression, given an input RGB ima ge. We rely on a recently introduced, expressive full body statistical 3d human model, GHUM, trained end-to-end, and learn to reconstruct its pose and shape sta te in a self-supervised regime. Central to our methodology, is a learning to lea rn and optimize approach, referred to as HUman Neural Descent (HUND), which avoi ds both second-order differentiation when training the model parameters, and exp ensive state gradient descen tin order to accurately minimize a semantic differe ntiable rendering loss at test time. Instead, we rely on novel recurrent stages to update the pose and shape parameters such that not only losses are minimized effectively, but the process is meta-regularized in order to ensure end-progress . HUND's symmetry between training and testing makes it the first 3d human sensi ng architecture to natively support different operating regimes including self-s upervised ones. In diverse tests, we show that HUND achieves very competitive re sults in datasets like H3.6M and 3DPW, as well as good quality 3d reconstruction s for complex imagery collected in-the-wild.
****
HR-NAS: Searching Efficient High-Resolution Neural Architectures With Lightweigh t Transformers
Mingyu Ding, Xiaochen Lian, Linjie Yang, Peng Wang, Xiaojie Jin, Zhiwu Lu, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2021, pp. 2982-2992
High-resolution representations (HR) are essential for dense prediction tasks su ch as segmentation, detection, and pose estimation. Learning HR representations is typically ignored in previous Neural Architecture Search (NAS) methods that f ocus on image classification. This work proposes a novel NAS method, called HR-N AS, which is able to find efficient and accurate networks for different tasks, b y effectively encoding multiscale contextual information while maintaining high-

resolution representations. In HR-NAS, we renovate the NAS search space as well as its searching strategy. To better encode multiscale image contexts in the search space of HR-NAS, we first carefully design a lightweight transformer, whose computational complexity can be dynamically changed with respect to different objective functions and computation budgets. To maintain high-resolution representations of the learned networks, HR-NAS adopts a multi-branch architecture that provides convolutional encoding of multiple feature resolutions, inspired by HRNet. Last, we proposed an efficient fine-grained search strategy to train HR-NAS, which effectively explores the search space, and finds optimal architectures given various tasks and computation resources. HR-NAS is capable of achieving state-of-the-art trade-offs between performance and FLOPs for three dense prediction tasks and an image classification task, given only small computational budgets. For example, HR-NAS surpasses SqueezeNAS that is specially designed for semantic segmentation by a large margin of 3.61% while improving efficiency by 45.9%. Code is available at https://github.com/dingmyu/HR-NAS.
****
Transitional Adaptation of Pretrained Models for Visual Storytelling
Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, Gunhee Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12658-12668
Previous models for vision-to-language generation tasks usually pretrain a visual encoder and a language generator in the respective domains and jointly finetune them with the target task. However, this direct transfer practice may suffer from the discord between visual specificity and language fluency since they are often separately trained from large corpora of visual and text data with no common ground. In this work, we claim that a transitional adaptation task is required between pretraining and finetuning to harmonize the visual encoder and the language model for challenging downstream target tasks like visual storytelling. We propose a novel approach named Transitional Adaptation of Pretrained Model (TAPM) that adapts the multi-modal modules to each other with a simpler alignment task between visual inputs only with no need for text labels. Through extensive experiments, we show that the adaptation step significantly improves the performance of multiple language models for sequential video and image captioning tasks. We achieve new state-of-the-art performance on both language metrics and human evaluation in the multi-sentence description task of LSMDC 2019 and the image storytelling task of VIST. Our experiments reveal that this improvement in caption quality does not depend on the specific choice of language models.
****
Improving Panoptic Segmentation at All Scales
Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7302-7311
Crop-based training strategies decouple training resolution from GPU memory consumption, allowing the use of large-capacity panoptic segmentation networks on multi-megapixel images. Using crops, however, can introduce a bias towards truncating or missing large objects. To address this, we propose a novel crop-aware bounding box regression loss (CABB loss), which promotes predictions to be consistent with the visible parts of the cropped objects, while not over-penalizing them for extending outside of the crop. We further introduce a novel data sampling and augmentation strategy which improves generalization across scales by counteracting the imbalanced distribution of object sizes. Combining these two contributions with a carefully designed, top-down panoptic segmentation architecture, we obtain new state-of-the-art results on the challenging Mapillary Vistas (MVD), Indian Driving and Cityscapes datasets, surpassing the previously best approach on MVD by +4.5% PQ and +5.2% mAP.
****
Model-Contrastive Federated Learning
Qinbin Li, Bingsheng He, Dawn Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10713-10722
Federated learning enables multiple parties to collaboratively train a machine l

earning model without communicating their local data. A key challenge in federat
ed learning is to handle the heterogeneity of local data distribution across par
ties. Although many studies have been proposed to address this challenge, we fin
d that they fail to achieve high performance in image datasets with deep learnin
g models. In this paper, we propose MOON: model-contrastive federated learning.
MOON is a simple and effective federated learning framework. The key idea of MOO
N is to utilize the similarity between model representations to correct the loca
l training of individual parties, i.e., conducting contrastive learning in model
-level. Our extensive experiments show that MOON significantly outperforms the o
ther state-of-the-art federated learning algorithms on various image classificat
ion tasks.
****
Scalability vs. Utility: Do We Have To Sacrifice One for the Other in Data Impor
tance Quantification?
Ruoxi Jia, Fan Wu, Xuehui Sun, Jiacen Xu, David Dao, Bhavya Kailkhura, Ce Zhang,
 Bo Li, Dawn Song; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2021, pp. 8239-8247
Quantifying the importance of each training point to a learning task is a fundam
ental problem in machine learning and the estimated importance scores have been
leveraged to guide a range of data workflows such as data summarization and doma
in adaption. One simple idea is to use the leave-one-out error of each training
point to indicate its importance. Recent work has also proposed to use the Shapl
ey value, as it defines a unique value distribution scheme that satisfies a set
of appealing properties. However, calculating Shapley values is often expensive,
 which limits its applicability in real-world applications at scale. Multiple he
uristics to improve the scalability of calculating Shapley values have been prop
osed recently, with the potential risk of compromising their utility in real-wor
ld applications. How well do existing data quantification methods perform on exi
sting workflows? How do these methods compare with each other, empirically and t
heoretically? Must we sacrifice scalability for the utility in these workflows w
hen using these methods? In this paper, we conduct a novel theoretical analysis
comparing the utility of different importance quantification methods, and report
 extensive experimental studies on settings such as noisy label detection, water
mark removal, data summarization, data acquisition, and domain adaptation on exi
sting and proposed workflows. We show that Shapley value approximation based on
a KNN surrogate over pre-trained feature embeddings obtains comparable utility w
ith existing algorithms while achieving significant scalability improvement, oft
en by orders of magnitude. Our theoretical analysis also justifies its advantage
 over the leave-one-out error. The code is available at https://github.com/AI-se
cure/Shapley-Study.
****
Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Ass
essment
Dongyu She, Yu-Kun Lai, Gaoxiong Yi, Kun Xu; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8475-8484
Learning computational models of image aesthetics can have a substantial impact
on visual art and graphic design. Although automatic image aesthetics assessment
 is a challenging topic by its subjective nature, psychological studies have con
firmed a strong correlation between image layouts and perceived image quality. W
hile previous state-of-the-art methods attempt to learn holistic information usi
ng deep Convolutional Neural Networks (CNNs), our approach is motivated by the f
act that Graph Convolutional Network (GCN) architecture is conceivably more suit
ed for modeling complex relations among image regions than vanilla convolutional
 layers. Specifically, we present a Hierarchical Layout-Aware Graph Convolutiona
l Network (HLA-GCN) to capture layout information. It is a dedicated double-subn
et neural network consisting of two LAGCN modules. The first LA-GCN module const
ructs an aesthetics-related graph in the coordinate space and performs reasoning
 over spatial nodes. The second LA-GCN module performs graph reasoning after agg
regating significant regions in a latent space. The model output is a hierarchic
al representation with layout-aware features from both spatial and aggregated no

des for unified aesthetics assessment. Extensive evaluations show that our proposed model outperforms the state-of-the-art on the AVA and AADB datasets across three different tasks. The code is available at http://github.com/days1011/HLAGCN.
****
Normalized Avatar Synthesis Using StyleGAN and Perceptual Refinement
Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11662-11672
We introduce a highly robust GAN-based framework for digitizing a normalized 3D avatar of a person from a single unconstrained photo. While the input image can be of a smiling person or taken in extreme lighting conditions, our method can reliably produce a high-quality textured model of a person's face in neutral expression and skin textures under diffuse lighting condition. Cutting-edge 3D face reconstruction methods use non-linear morphable face models combined with GAN-based decoders to capture the likeness and details of a person but fail to produce neutral head models with unshaded albedo textures which is critical for creating relightable and animation-friendly avatars for integration in virtual environments. The key challenges for existing methods to work is the lack of training and ground truth data containing normalized 3D faces. We propose a two-stage approach to address this problem. First, we adopt a highly robust normalized 3D face generator by embedding a non-linear morphable face model into a StyleGAN2 network. This allows us to generate detailed but normalized facial assets. This inference is then followed by a perceptual refinement step that uses the generated assets as regularization to cope with the limited available training samples of normalized faces. We further introduce a Normalized Face Dataset, which consists of a combination photogrammetry scans, carefully selected photographs, and generated fake people with neutral expressions in diffuse lighting conditions. While our prepared dataset contains two orders of magnitude less subjects than cutting edge GAN-based 3D facial reconstruction methods, we show that it is possible to produce high-quality normalized face models for very challenging unconstrained input images, and demonstrate superior performance to the current state-of-the-art.
****
CT-Net: Complementary Transfering Network for Garment Transfer With Arbitrary Geometric Changes
Fan Yang, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9899-9908
Garment transfer shows great potential in realistic applications with the goal of transfering outfits across different people images. However, garment transfer between images with heavy misalignments or severe occlusions still remains as a challenge. In this work, we propose Complementary Transfering Network (CT-Net) to adaptively model different levels of geometric changes and transfer outfits between different people. In specific, CT-Net consists of three modules: i) A complementary warping module first estimates two complementary warpings to transfer the desired clothes in different granularities. ii) A layout prediction module is proposed to predict the target layout, which guides the preservation or generation of the body parts in the synthesized images. iii) A dynamic fusion module adaptively combines the advantages of the complementary warpings to render the garment transfer results. Extensive experiments conducted on DeepFashion dataset demonstrate that our network synthesizes high-quality garment transfer images and significantly outperforms the state-of-art methods both qualitatively and quantitatively. Our source code will be available online.
****
MetaCorrection: Domain-Aware Meta Loss Correction for Unsupervised Domain Adaptation in Semantic Segmentation
Xiaoqing Guo, Chen Yang, Baopu Li, Yixuan Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3927-3936
Unsupervised domain adaptation (UDA) aims to transfer the knowledge from the labeled source domain to the unlabeled target domain. Existing self-training based

UDA approaches assign pseudo labels for target data and treat them as ground truth labels to fully leverage unlabeled target data for model adaptation. However, the generated pseudo labels from the model optimized on the source domain inevitably contain noise due to the domain gap. To tackle this issue, we advance a MetaCorrection framework, where a Domain-aware Meta-learning strategy is devised to benefit Loss Correction (DMLC) for UDA semantic segmentation. In particular, we model the noise distribution of pseudo labels in target domain by introducing a noise transition matrix (NTM) and construct meta data set with domain-invariant source data to guide the estimation of NTM. Through the risk minimization on the meta data set, the optimized NTM thus can correct the noisy issues in pseudo labels and enhance the generalization ability of the model on the target data. Considering the capacity gap between shallow and deep features, we further employ the proposed DMLC strategy to provide matched and compatible supervision signals for different level features, thereby ensuring deep adaptation. Extensive experimental results highlight the effectiveness of our method against existing state-of-the-art methods on three benchmarks.
****
Multi-Stage Progressive Image Restoration
Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14821-14831
Image restoration tasks demand a complex balance between spatial details and high-level contextualized information while recovering images. In this paper, we propose a novel synergistic design that can optimally balance these competing goals. Our main proposal is a multi-stage architecture, that progressively learns restoration functions for the degraded inputs, thereby breaking down the overall recovery process into more manageable steps. Specifically, our model first learns the contextualized features using encoder-decoder architectures and later combines them with a high-resolution branch that retains local information. At each stage, we introduce a novel per-pixel adaptive design that leverages in-situ supervised attention to reweight the local features. A key ingredient in such a multi-stage architecture is the information exchange between different stages. To this end, we propose a two-faceted approach where the information is not only exchanged sequentially from early to late stages, but lateral connections between feature processing blocks also exist to avoid any loss of information. The resulting tightly interlinked multi-stage architecture, named as MPRNet, delivers strong performance gains on ten datasets across a range of tasks including image deraining, deblurring, and denoising. The source code and pre-trained models are available at https://github.com/swz30/MPRNet.
****
PointNetLK Revisited
Xueqian Li, Jhony Kaesemodel Pontes, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12763-12772
We address the generalization ability of recent learning-based point cloud registration methods. Despite their success, these approaches tend to have poor performance when applied to mismatched conditions that are not well-represented in the training set, such as unseen object categories, different complex scenes, or unknown depth sensors. In these circumstances, it has often been better to rely on classical non-learning methods (e.g., Iterative Closest Point), which have better generalization ability. Hybrid learning methods, that use learning for predicting point correspondences and then a deterministic step for alignment, have offered some respite, but are still limited in their generalization abilities. We revisit a recent innovation---PointNetLK---and show that the inclusion of an analytical Jacobian can exhibit remarkable generalization properties while reaping the inherent fidelity benefits of a learning framework. Our approach not only outperforms the state-of-the-art in mismatched conditions but also produces results competitive with current learning methods when operating on real-world test data close to the training set.
****

Deep Convolutional Dictionary Learning for Image Denoising
Hongyi Zheng, Hongwei Yong, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 630-641

Inspired by the great success of deep neural networks (DNNs), many unfolding methods have been proposed to integrate traditional image modeling techniques, such as dictionary learning (DicL) and sparse coding, into DNNs for image restoration. However, the performance of such methods remains limited for several reasons. First, the unfolded architectures do not strictly follow the image representation model of DicL and lose the desired physical meaning. Second, handcrafted priors are still used in most unfolding methods without effectively utilizing the learning capability of DNNs. Third, a universal dictionary is learned to represent all images, reducing the model representation flexibility. We propose a novel framework of deep convolutional dictionary learning (DCDicL), which follows the representation model of DicL strictly, learns the priors for both representation coefficients and the dictionaries, and can adaptively adjust the dictionary for each input image based on its content. The effectiveness of our DCDicL method is validated on the image denoising problem. DCDicL demonstrates leading denoising performance in terms of both quantitative metrics (e.g., PSNR, SSIM) and visual quality. In particular, it can reproduce the subtle image structures and textures, which are hard to recover by many existing denoising DNNs. The code is available at: https://github.com/natezhenghy/DCDicL_denoising.

****

Fourier Contour Embedding for Arbitrary-Shaped Text Detection
Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, Wayne Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3123-3131

One of the main challenges for arbitrary-shaped text detection is to design a good text instance representation that allows networks to learn diverse text geometry variances. Most of existing methods model text instances in image spatial domain via masks or contour point sequences in the Cartesian or the polar coordinate system. However, the mask representation might lead to expensive post-processing, while the point sequence one may have limited capability to model texts with highly-curved shapes. To tackle these problems, we model text instances in the Fourier domain and propose one novel Fourier Contour Embedding (FCE) method to represent arbitrary shaped text contours as compact signatures. We further construct FCENet with a backbone, feature pyramid networks (FPN) and a simple post-processing with the Inverse Fourier Transformation (IFT) and Non-Maximum Suppression (NMS). Different from previous methods, FCENet first predicts compact Fourier signatures of text instances, and then reconstructs text contours via IFT and NMS during test. Extensive experiments demonstrate that FCE is accurate and robust to fit contours of scene texts even with highly-curved shapes, and also validate the effectiveness and the good generalization of FCENet for arbitrary-shaped text detection. Furthermore, experimental results show that our FCENet is superior to the state-of-the-art (SOTA) methods on CTW1500 and Total-Text, especially on challenging highly-curved text subset.

****

TAP: Text-Aware Pre-Training for Text-VQA and Text-Caption
Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8751-8761

In this paper, we propose Text-Aware Pre-training (TAP) for Text-VQA and Text-Caption tasks. These two tasks aim at reading and understanding scene text in images for question answering and image caption generation, respectively. In contrast to the conventional vision-language pre-training that fails to capture scene text and its relationship with the visual and text modalities, TAP explicitly incorporates scene text (generated from OCR engines) in pre-training. With three pre-training tasks, including masked language modeling (MLM), image-text (contrastive) matching (ITM), and relative (spatial) position prediction (RPP), TAP effectively helps the model learn a better aligned representation among the three modalities: text word, visual object, and scene text. Due to this aligned represent

ation learning, even pre-trained on the same downstream task dataset, TAP alread
y boosts the absolute accuracy on the TextVQA dataset by +5.4%, compared with a
non-TAP baseline. To further improve the performance, we build a large-scale dat
aset based on the Conceptual Caption dataset, named OCR-CC, which contains 1.4 m
illion scene text-related image-text pairs. Pre-trained on this OCR-CC dataset,
our approach outperforms the state of the art by large margins on multiple tasks
, i.e., +8.3% accuracy on TextVQA, +8.6% accuracy on ST-VQA, and +10.2 CIDEr sco
re on TextCaps.
****
Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representatio
n Learning
Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, Jianlong Fu; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2021, pp. 12976-12985
We study on joint learning of Convolutional Neural Network (CNN) and Transformer
 for vision-language pre-training (VLPT) which aims to learn cross-modal alignme
nts from millions of image-text pairs. State-of-the-art approaches extract salie
nt image regions and align regions with words step-by-step. As region-based repr
esentations usually represent parts of an image, it is challenging for existing
models to fully understand the semantics from paired natural languages. In this
paper, we propose SOHO to ""See Out of tHe bOx"" that takes a full image as inpu
t, and learns vision-language representation in an end-to-end manner. SOHO does
not require bounding box annotations, while enables 10 times faster inference th
an region-based approaches. In particular, SOHO learns to extract comprehensive
yet compact image features through a visual dictionary (VD) that facilitates cro
ss-modal understanding. VD is designed to represent consistent visual abstractio
ns of similar semantics, and VD can be further updated on-the-fly during pre-tra
ining. We conduct experiments on four well-established vision-language tasks by
following standard VLPT settings. SOHO achieves absolute gains of 2.0% R@1 score
 on MSCOCO text retrieval 5k test split, 1.5% accuracy on NLVR2 test-P split, 6.
7% accuracy on SNLI-VE test split, respectively.
****
Quality-Agnostic Image Recognition via Invertible Decoder
Insoo Kim, Seungju Han, Ji-won Baek, Seong-Jin Park, Jae-Joon Han, Jinwoo Shin;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2021, pp. 12257-12266
Despite the remarkable performance of deep models on image recognition tasks, th
ey are known to be susceptible to common corruptions such as blur, noise, and lo
w-resolution. Data augmentation is a conventional way to build a robust model by
 considering these common corruptions during the training. However, a naive data
 augmentation scheme may result in a non-specialized model for particular corrup
tions, as the model tends to learn the averaged distribution among corruptions.
To mitigate the issue, we propose a new paradigm of training deep image recognit
ion networks that produce clean-like features from any quality image via an inve
rtible neural architecture. The proposed method consists of two stages. In the f
irst stage, we train an invertible network with only clean images under the reco
gnition objective. In the second stage, its inversion, i.e., the invertible deco
der, is attached to a new recognition network and we train this encoder-decoder
network using both clean and corrupted images by considering recognition and rec
onstruction objectives. Our two-stage scheme allows the network to produce clean
-like and robust features from any quality images, by reconstructing their clean
 images via the invertible decoder. We demonstrate the effectiveness of our meth
od on image classification and face recognition tasks.
****
Hybrid Rotation Averaging: A Fast and Robust Rotation Averaging Approach
Yu Chen, Ji Zhao, Laurent Kneip; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2021, pp. 10358-10367
We address rotation averaging (RA) and its application to real-world 3D reconstr
uction. Local optimisation based approaches are the de facto choice, though they
 only guarantee a local optimum. Global optimisers ensure global optimality in l

ow noise conditions, but they are inefficient and may easily deviate under the influence of outliers or elevated noise levels. We push the envelope of rotation averaging by leveraging the advantages of a global RA method and a local RA method. Combined with a fast view graph filtering as preprocessing, the proposed hybrid approach is robust to outliers. We further apply the proposed hybrid rotation averaging approach to incremental Structure from Motion (SfM), the accuracy and robustness of SfM are both improved by adding the resulting global rotations as regularisers to bundle adjustment. Overall, we demonstrate high practicality of the proposed method as bad camera poses are effectively corrected and drift is reduced.
****
One Thing One Click: A Self-Training Approach for Weakly Supervised 3D Semantic Segmentation
Zhengzhe Liu, Xiaojuan Qi, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1726-1736
Point cloud semantic segmentation often requires largescale annotated training data, but clearly, point-wise labels are too tedious to prepare. While some recent methods propose to train a 3D network with small percentages of point labels, we take the approach to an extreme and propose "One Thing One Click," meaning that the annotator only needs to label one point per object. To leverage these extremely sparse labels in network training, we design a novel self-training approach, in which we iteratively conduct the training and label propagation, facilitated by a graph propagation module. Also, we adopt a relation network to generate per-category prototype and explicitly model the similarity among graph nodes to generate pseudo labels to guide the iterative training. Experimental results on both ScanNet-v2 and S3DIS show that our self-training approach, with extremely-sparse annotations, outperforms all existing weakly supervised methods for 3D semantic segmentation by a large margin, and our results are also comparable to those of the fully supervised counterparts.
****
Out-of-Distribution Detection Using Union of 1-Dimensional Subspaces
Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9452-9461
The goal of out-of-distribution (OOD) detection is to handle the situations where the test samples are drawn from a different distribution than the training data. In this paper, we argue that OOD samples can be detected more easily if the training data is embedded into a low-dimensional space, such that the embedded training samples lie on a union of 1-dimensional subspaces. We show that such embedding of the in-distribution (ID) samples provides us with two main advantages. First, due to compact representation in the feature space, OOD samples are less likely to occupy the same region as the known classes. Second, the first singular vector of ID samples belonging to a 1-dimensional subspace can be used as their robust representative. Motivated by these observations, we train a deep neural network such that the ID samples are embedded onto a union of 1-dimensional subspaces. At the test time, employing sampling techniques used for approximate Bayesian inference in deep learning, input samples are detected as OOD if they occupy the region corresponding to the ID samples with probability 0. Spectral components of the ID samples are used as robust representative of this region. Our method does not have any hyperparameter to be tuned using extra information and it can be applied on different modalities with minimal change. The effectiveness of the proposed method is demonstrated on different benchmark datasets, both in the image and video classification domains.
****
MP3: A Unified Model To Map, Perceive, Predict and Plan
Sergio Casas, Abbas Sadat, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14403-14412
High-definition maps (HD maps) are a key component of most modern self-driving systems due to their valuable semantic and geometric information. Unfortunately, building HD maps has proven hard to scale due to their cost as well as the requi

rements they impose in the localization system that has to work everywhere with centimeter-level accuracy. Being able to drive without an HD map would be very beneficial to scale self-driving solutions as well as to increase the failure tolerance of existing ones (e.g., if localization fails or the map is not up-to-date). Towards this goal, we propose an end-to-end approach to mapless driving where the input is raw sensor data and a high-level command (e.g., turn left at the intersection). We then predict intermediate representations in the form of an online map and the current and future state of dynamic agents, and exploit them in a novel neural motion planner to make interpretable decisions taking into account uncertainty. We show that our approach is significantly safer, more comfortable, and can follow commands better than the baselines in challenging long-term closed-loop simulations, as well as when compared to an expert driver in a large-scale real-world dataset.

****

SCALE: Modeling Clothed Humans with a Surface Codec of Articulated Local Elements

Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16082-16093

Learning to model and reconstruct humans in clothing is challenging due to articulation, non-rigid deformation, and varying clothing types and topologies. To enable learning, the choice of representation is the key. Recent work uses neural networks to parameterize local surface elements. This approach captures locally coherent geometry and non-planar details, can deal with varying topology, and does not require registered training data. However, naively using such methods to model 3D clothed humans fails to capture fine-grained local deformations and generalizes poorly. To address this, we present three key innovations: First, we deform surface elements based on a human body model such that large-scale deformations caused by articulation are explicitly separated from topological changes and local clothing deformations. Second, we address the limitations of existing neural surface elements by regressing local geometry from local features, significantly improving the expressiveness. Third, we learn a pose embedding on a 2D parameterization space that encodes posed body geometry, improving generalization to unseen poses by reducing non-local spurious correlations. We demonstrate the efficacy of our surface representation by learning models of complex clothing from point clouds. The clothing can change topology and deviate from the topology of the body. Once learned, we can animate previously unseen motions, producing high-quality point clouds, from which we generate realistic images with neural rendering. We assess the importance of each technical contribution and show that our approach outperforms the state-of-the-art methods in terms of reconstruction accuracy and inference time. The code is available for research purposes at https://qianlim.github.io/SCALE.

****

Playable Video Generation

Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10061-10070

This paper introduces the unsupervised learning problem of playable video generation (PVG). In PVG, we aim at allowing a user to control the generated video by selecting a discrete action at every time step as when playing a video game. The difficulty of the task lies both in learning semantically consistent actions and in generating realistic videos conditioned on the user input. We propose a novel framework for PVG that is trained in a self-supervised manner on a large dataset of unlabelled videos. We employ an encoder-decoder architecture where the predicted action labels act as bottleneck. The network is constrained to learn a rich action space using, as main driving loss, a reconstruction loss on the generated video. We demonstrate the effectiveness of the proposed approach on several datasets with wide environment variety. Further details, code and examples are available on our project page willi-menapace.github.io/playable-video-generation-website.

****
AdCo: Adversarial Contrast for Efficient Learning of Unsupervised Representation
s From Self-Trained Negative Adversaries
Qianjiang Hu, Xiao Wang, Wei Hu, Guo-Jun Qi; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1074-1083
Contrastive learning relies on constructing a collection of negative examples th
at are sufficiently hard to discriminate against positive queries when their rep
resentations are self-trained. Existing contrastive learning methods either main
tain a queue of negative samples over mini-batches while only a small portion of
 them are updated in an iteration or only use the other examples from the curren
t minibatch as negatives. They could not closely track the change of the learned
 representation over iterations by updating the entire queue as a whole or disca
rd the useful information from the past mini-batches. Alternatively, we present
to directly learn a set of negative adversaries playing against the self-trained
 representation. Two players, the representation network and negative adversarie
s are alternately updated to obtain the most challenging negative examples again
st which the representation of positive queries will be trained to discriminate.
 We further show that the negative adversaries are updated towards a weighted co
mbination of positive queries by maximizing the adversarial contrastive loss, th
ereby allowing them to closely track the change of representations over time. Ex
periment results demonstrate the proposed Adversarial Contrastive (AdCo) model n
ot only achieves superior performances (a top-1 accuracy of 73.2% over 200 epoch
s and 75.7% over 800 epochs with linear evaluation on ImageNet), but also can be
 pre-trained more efficiently with much shorter GPU time and fewer epochs. The s
ource code is available at https://github.com/maple-research-lab/AdCo.
****
Permute, Quantize, and Fine-Tune: Efficient Compression of Neural Networks
Julieta Martinez, Jashan Shewakramani, Ting Wei Liu, Ioan Andrei Barsan, Wenyuan
 Zeng, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2021, pp. 15699-15708
Compressing large neural networks is an important step for their deployment in r
esource-constrained computational platforms. In this context, vector quantizatio
n is an appealing framework that expresses multiple parameters using a single co
de, and has recently achieved state-of-the-art network compression on a range of
 core vision and natural language processing tasks. Key to the success of vector
 quantization is deciding which parameter groups should be compressed together.
Previous work has relied on heuristics that group the spatial dimension of indiv
idual convolutional filters, but a general solution remains unaddressed. This is
 desirable for pointwise convolutions (which dominate modern architectures), lin
ear layers (which have no notion of spatial dimension), and convolutions (when m
ore than one filter is compressed to the same codeword). In this paper we make t
he observation that the weights of two adjacent layers can be permuted while exp
ressing the same function. We then establish a connection to rate-distortion the
ory and search for permutations that result in networks that are easier to compr
ess. Finally, we rely on an annealed quantization algorithm to better compress t
he network and achieve higher final accuracy. We show results on image classific
ation, object detection, and segmentation, reducing the gap with the uncompresse
d model by 40 to 70% w.r.t. the current state of the art. We will release code t
o reproduce all our experiments.
****
Mol2Image: Improved Conditional Flow Models for Molecule to Image Synthesis
Karren Yang, Samuel Goldman, Wengong Jin, Alex X. Lu, Regina Barzilay, Tommi Jaa
kkola, Caroline Uhler; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2021, pp. 6688-6698
In this paper, we aim to synthesize cell microscopy images under different molec
ular interventions, motivated by practical applications to drug development. Bui
lding on the recent success of graph neural networks for learning molecular embe
ddings and flow-based models for image generation, we propose Mol2Image: a flow-
based generative model for molecule to cell image synthesis. To generate cell fe
atures at different resolutions and scale to high-resolution images, we develop

a novel multi-scale flow architecture based on a Haar wavelet image pyramid. To maximize the mutual information between the generated images and the molecular interventions, we devise a training strategy based on contrastive learning. To evaluate our model, we propose a new set of metrics for biological image generation that are robust, interpretable, and relevant to practitioners. We show quantitatively that our method learns a meaningful embedding of the molecular intervention, which is translated into an image representation reflecting the biological effects of the intervention.
****
Improved Handling of Motion Blur in Online Object Detection
Mohamed Sayed, Gabriel Brostow; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1706-1716
We wish to detect specific categories of objects, for online vision systems that will run in the real world. Object detection is already very challenging. It is even harder when the images are blurred, from the camera being in a car or a hand-held phone. Most existing efforts either focused on sharp images, with easy to label ground truth, or they have treated motion blur as one of many generic corruptions. Instead, we focus especially on the details of egomotion induced blur. We explore five classes of remedies, where each targets different potential causes for the performance gap between sharp and blurred images. For example, first deblurring an image changes its human interpretability, but at present, only partly improves object detection. The other four classes of remedies address multi-scale texture, out-of-distribution testing, label generation, and conditioning by blur-type. Surprisingly, we discover that custom label generation aimed at resolving spatial ambiguity, ahead of all others, markedly improves object detection. Also, in contrast to findings from classification, we see a noteworthy boost by conditioning our model on bespoke categories of motion blur. We validate and cross-breed the different remedies experimentally on blurred COCO images and real-world blur datasets, producing an easy and practical favorite model with superior detection rates.
****
Multimodal Motion Prediction With Stacked Transformers
Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7577-7586
Predicting multiple plausible future trajectories of the nearby vehicles is crucial for the safety of autonomous driving. Recent motion prediction approaches attempt to achieve such multimodal motion prediction by implicitly regularizing the feature or explicitly generating multiple candidate proposals. However, it remains challenging since the latent features may concentrate on the most frequent mode of the data while the proposal-based methods depend largely on the prior knowledge to generate and select the proposals. In this work, we propose a novel transformer framework for multimodal motion prediction, termed as mmTransformer. A novel network architecture based on stacked transformers is designed to model the multimodality at feature level with a set of fixed independent proposals. A region-based training strategy is then developed to induce the multimodality of the generated proposals. Experiments on Argoverse dataset show that the proposed model achieves the state-of-the-art performance on motion prediction, substantially improving the diversity and the accuracy of the predicted trajectories. Demo video and code are available at https://decisionforce. github.io/mmTransformer.
****
The Translucent Patch: A Physical and Universal Attack on Object Detectors
Alon Zolfi, Moshe Kravchik, Yuval Elovici, Asaf Shabtai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15232-15241
Physical adversarial attacks against object detectors have seen increasing success in recent years. However, these attacks require direct access to the object of interest in order to apply a physical patch. Furthermore, to hide multiple objects, an adversarial patch must be applied to each object. In this paper, we pro

pose a contactless translucent physical patch containing a carefully constructed pattern, which is placed on the camera's lens, to fool state-of-the-art object detectors. The primary goal of our patch is to hide all instances of a selected target class. In addition, the optimization method used to construct the patch aims to ensure that the detection of other (untargeted) classes remains unharmed. Therefore, in our experiments, which are conducted on state-of-the-art object detection models used in autonomous driving, we study the effect of the patch on the detection of both the selected target class and the other classes. We show that our patch was able to prevent the detection of 42.27% of all stop sign instances while maintaining high (nearly 80%) detection of the other classes.
****
Exploit Visual Dependency Relations for Semantic Segmentation
Mingyuan Liu, Dan Schonfeld, Wei Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9726-9735
Dependency relations among visual entities are ubiquity because both objects and scenes are highly structured. They provide prior knowledge about the real world that can help improve the generalization ability of deep learning approaches. Different from contextual reasoning which focuses on feature aggregation in the spatial domain, visual dependency reasoning explicitly models the dependency relations among visual entities. In this paper, we introduce a novel network architecture, termed the dependency network or DependencyNet, for semantic segmentation. It unifies dependency reasoning at three semantic levels. Intra-class reasoning decouples the representations of different object categories and updates them separately based on the internal object structures. Inter-class reasoning then performs spatial and semantic reasoning based on the dependency relations among different object categories. We will have an in-depth investigation on how to discover the dependency graph from the training annotations. Global dependency reasoning further refines the representations of each object category based on the global scene information. Extensive ablative studies with a controlled model size and the same network depth show that each individual dependency reasoning component benefits semantic segmentation and they together significantly improve the base network. Experimental results on two benchmark datasets show the DependencyNet achieves comparable performance to the recent states of the art.
****
Dense Label Encoding for Boundary Discontinuity Free Rotation Detection
Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, Junchi Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15819-15829
Rotation detection serves as a fundamental building block in many visual applications involving aerial image, scene text, and face etc. Differing from the dominant regression-based approaches for orientation estimation, this paper explores a relatively less-studied methodology based on classification. The hope is to inherently dismiss the boundary discontinuity issue as encountered by the regression-based detectors. We propose new techniques to push its frontier in two aspects: i) new encoding mechanism: the design of two Densely Coded Labels (DCL) for angle classification, to replace the Sparsely Coded Label (SCL) in existing classification-based detectors, leading to three times training speed increase as empirically observed across benchmarks, further with notable improvement in detection accuracy; ii) loss re-weighting: we propose Angle Distance and Aspect Ratio Sensitive Weighting (ADARSW), which improves the detection accuracy especially for square-like objects, by making DCL-based detectors sensitive to angular distance and object's aspect ratio. Extensive experiments and visual analysis on large-scale public datasets for aerial images i.e. DOTA, UCAS-AOD, HRSC2016, as well as scene text dataset ICDAR2015 and MLT, show the effectiveness of our approach. The source code will be made public available.
****
Self-Supervised Collision Handling via Generative 3D Garment Models for Virtual Try-On
Igor Santesteban, Nils Thuerey, Miguel A. Otaduy, Dan Casas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.

 11763-11773
We propose a new generative model for 3D garment deformations that enables us to
 learn, for first time, a data-driven method for virtual try-on that effectively
 addresses garment-body collisions. In contrast to existing methods that require
 an undesirable postprocessing step to fix garment-body interpenetrations at tes
t time, our approach directly outputs 3D garment configurations that do not coll
ide with the underlying body. Key to our success is a new canonical space for ga
rments that removes pose-and-shape deformations already captured by a new diffus
ed human body model, which extrapolates body surface properties such as skinning
 weights and blendshapes to any 3D point. We leverage this representation to tra
in a generative model with a novel self-supervised collision term that learns to
 reliably solve garment-body interpenetrations. We extensively evaluate and comp
are our results with recently proposed data-driven methods, and show that our me
thod is the first to successfully address garment-body contact in unseen body sh
apes and motions, without compromising the realism and detail.
****
DexYCB: A Benchmark for Capturing Hand Grasping of Objects
Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay
, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, Diete
r Fox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2021, pp. 9044-9053
We introduce DexYCB, a new dataset for capturing hand grasping of objects. We fi
rst compare DexYCB with a related one through cross-dataset evaluation. We then
present a thorough benchmark of state-of-the-art approaches on three relevant ta
sks: 2D object and keypoint detection, 6D object pose estimation, and 3D hand po
se estimation. Finally, we evaluate a new robotics-relevant task: generating saf
e robot grasps in human-to-robot object handover.
****
Prototype Completion With Primitive Knowledge for Few-Shot Learning
Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, Lisai Zhang; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
 pp. 3754-3762
Few-shot learning is a challenging task, which aims to learn a classifier for no
vel classes with few examples. Pre-training based meta-learning methods effectiv
ely tackle the problem by pre-training a feature extractor and then fine-tuning
it through the nearest centroid based meta-learning. However, results show that
the fine-tuning step makes very marginal improvements. In this paper, 1) we figu
re out the key reason, i.e., in the pre-trained feature space, the base classes
already form compact clusters while novel classes spread as groups with large va
riances, which implies that fine-tuning the feature extractor is less meaningful
; 2) instead of fine-tuning the feature extractor, we focus on estimating more r
epresentative prototypes during meta-learning. Consequently, we propose a novel
prototype completion based meta-learning framework. This framework first introdu
ces primitive knowledge (i.e., class-level part or attribute annotations) and ex
tracts representative attribute features as priors. Then, we design a prototype
completion network to learn to complete prototypes with these priors. To avoid t
he prototype completion error caused by primitive knowledge noises or class diff
erences, we further develop a Gaussian based prototype fusion strategy that comb
ines the mean-based and completed prototypes by exploiting the unlabeled samples
. Extensive experiments show that our method: (i) can obtain more accurate proto
types; (ii) outperforms state-of-the-art techniques by 2%   9% in terms of class
ification accuracy. Our code is available online.
****
High-Quality Stereo Image Restoration From Double Refraction
Hakyeong Kim, Andreas Meuleman, Daniel S. Jeon, Min H. Kim; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
11987-11995
Single-shot monocular birefractive stereo methods have been used for estimating
sparse depth from double refraction over edges. They also obtain an ordinary-ray
 (o-ray) image concurrently or subsequently through additional post-processing o

f depth densification and deconvolution. However, when an extraordinary-ray (e-ray) image is restored to acquire stereo images, the existing methods suffer from very severe restoration artifacts in stereo images due to a low signal-to-noise ratio of input e-ray image or depth/deconvolution errors. In this work, we present a novel stereo image restoration network that can restore stereo images directly from a double-refraction image. First, we built a physically faithful birefractive stereo imaging dataset by simulating the double refraction phenomenon with existing RGB-D datasets. Second, we formulated a joint stereo restoration problem that accounts for not only geometric relation between o-/e-ray images but also joint optimization of restoring both stereo images. We trained our model with our birefractive image dataset in an end-to-end manner. Our model restores high-quality stereo images directly from double refraction in real-time, enabling high-quality stereo video using a monocular camera. Our method also allows us to estimate dense depth maps from stereo images using a conventional stereo method. We evaluate the performance of our method experimentally and synthetically with the ground truth. Results validate that our stereo image restoration network outperforms the existing methods with high accuracy. We demonstrate several image-editing applications using our high-quality stereo images and dense depth maps.
****

Track, Check, Repeat: An EM Approach to Unsupervised Tracking

Adam W. Harley, Yiming Zuo, Jing Wen, Ayush Mangal, Shubhankar Potdar, Ritwick Chaudhry, Katerina Fragkiadaki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16581-16591

We propose an unsupervised method for detecting and tracking moving objects in 3D, in unlabelled RGB-D videos. The method begins with classic handcrafted techniques for segmenting objects using motion cues: we estimate optical flow and camera motion, and conservatively segment regions that appear to be moving independently of the background. Treating these initial segments as pseudo-labels, we learn an ensemble of appearance-based 2D and 3D detectors, under heavy data augmentation. We use this ensemble to detect new instances of the "moving" type, even if they are not moving, and add these as new pseudo-labels. Our method is an expectation-maximization algorithm, where in the expectation step we fire all modules and look for agreement among them, and in the maximization step we re-train the modules to improve this agreement. The constraint of ensemble agreement helps combat contamination of the generated pseudo-labels (during the E step), and data augmentation helps the modules generalize to yet-unlabelled data (during the M step). We compare against existing unsupervised object discovery and tracking methods, using challenging videos from CATER and KITTI, and show strong improvements over the state-of-the-art.
****

LayoutTransformer: Scene Layout Generation With Conceptual and Spatial Diversity

Cheng-Fu Yang, Wan-Cyuan Fan, Fu-En Yang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3732-3741

When translating text inputs into layouts or images, existing works typically require explicit descriptions of each object in a scene, including their spatial information or the associated relationships. To better exploit the text input, so that implicit objects or relationships can be properly inferred during layout generation, we propose a LayoutTransformer Network (LT-Net) in this paper. Given a scene-graph input, our LT-Net uniquely encodes the semantic features for exploiting their co-occurrences and implicit relationships. This allows one to manipulate conceptually diverse yet plausible layout outputs. Moreover, the decoder of our LT-Net translates the encoded contextual features into bounding boxes with self-supervised relation consistency preserved. By fitting their distributions to Gaussian mixture models, spatially-diverse layouts can be additionally produced by LT-Net. We conduct extensive experiments on the datasets of MS-COCO and Visual Genome, and confirm the effectiveness and plausibility of our LT-Net over recent layout generation models.
****

Practical Wide-Angle Portraits Correction With Deep Structured Models

Jing Tan, Shan Zhao, Pengfei Xiong, Jiangyu Liu, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3498-3506

Wide-angle portraits often enjoy expanded views. However, they contain perspective distortions, especially noticeable when capturing group portrait photos, where the background is skewed and faces are stretched. This paper introduces the first deep learning based approach to remove such artifacts from freely-shot photos. Specifically, given a wide-angle portrait as input, we build a cascaded network consisting of a LineNet, a ShapeNet, and a transition module (TM), which corrects perspective distortions on the background, adapts to the stereographic projection on facial regions, and achieves smooth transitions between these two projections, accordingly. To train our network, we build the first perspective portrait dataset with a large diversity in identities, scenes and camera modules. For the quantitative evaluation, we introduce two novel metrics, line consistency and face congruence. Compared to the previous state-of-the-art approach, our method does not require camera distortion parameters. We demonstrate that our approach significantly outperforms the previous state-of-the-art approach both qualitatively and quantitatively.

****

CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild

Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, Bodo Rosenhahn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13294-13304

Human pose estimation from single images is a challenging problem in computer vision that requires large amounts of labeled training data to be solved accurately. Unfortunately, for many human activities (e.g. outdoor sports) such training data does not exist and is hard or even impossible to acquire with traditional motion capture systems. We propose a self-supervised approach that learns a single image 3D pose estimator from unlabeled multi-view data. To this end, we exploit multi-view consistency constraints to disentangle the observed 2D pose into the underlying 3D pose and camera rotation. In contrast to most existing methods, we do not require calibrated cameras and can therefore learn from moving cameras. Nevertheless, in the case of a static camera setup, we present an optional extension to include constant relative camera rotations over multiple views into our framework. Key to the success are new, unbiased reconstruction objectives that mix information across views and training samples. The proposed approach is evaluated on two benchmark datasets (Human3.6M and MPII-INF-3DHP) and on the in-the-wild SkiPose dataset.

****

Pushing It Out of the Way: Interactive Visual Navigation

Kuo-Hao Zeng, Luca Weihs, Ali Farhadi, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9868-9877

We have observed significant progress in visual navigation for embodied agents. A common assumption in studying visual navigation is that the environments are static; this is a limiting assumption. Intelligent navigation may involve interacting with the environment beyond just moving forward/backward and turning left/right. Sometimes, the best way to navigate is to push something out of the way. In this paper, we study the problem of interactive navigation where agents learn to change the environment to navigate more efficiently to their goals. To this end, we introduce the Neural Interaction Engine (NIE) to explicitly predict the change in the environment caused by the agent's actions. By modeling the changes while planning, we find that agents exhibit significant improvements in their navigational capabilities. More specifically, we consider two downstream tasks in the physics-enabled, visually rich, AI2-THOR environment: (1) reaching a target while the path to the target is blocked (2) moving an object to a target location by pushing it. For both tasks, agents equipped with an NIE significantly outperform agents without the understanding of the effect of the actions indicating the benefits of our approach. The code and dataset are available at github.com/KuoHaoZeng/Interactive_Visual_Navigation.

****
Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillati
on
Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, Dong Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14090-14100
Weakly supervised phrase grounding aims at learning region-phrase correspondences using only image-sentence pairs. A major challenge thus lies in the missing links between image regions and sentence phrases during training. To address this challenge, we leverage a generic object detector at training time, and propose a contrastive learning framework that accounts for both region-phrase and image-sentence matching. Our core innovation is the learning of a region-phrase score function, based on which an image-sentence score function is further constructed. Importantly, our region-phrase score function is learned by distilling from soft matching scores between the detected object names and candidate phrases within an image-sentence pair, while the image-sentence score function is supervised by ground-truth image-sentence pairs. The design of such score functions removes the need of object detection at test time, thereby significantly reducing the inference cost. Without bells and whistles, our approach achieves state-of-the-art results on visual phrase grounding, surpassing previous methods that require expensive object detectors at test time.
****
EvDistill: Asynchronous Events To End-Task Learning via Bidirectional Reconstruc
tion-Guided Cross-Modal Knowledge Distillation
Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 608-619
Event cameras sense per-pixel intensity changes and produce asynchronous event streams with high dynamic range and less motion blur, showing advantages over the conventional cameras. A hurdle of training event-based models is the lack of large qualitative labeled data. Prior works learning end-tasks mostly rely on labeled or pseudo-labeled datasets obtained from the active pixel sensor (APS) frames; however, such datasets' quality is far from rivaling those based on the canonical images. In this paper, we propose a novel approach, called EvDistill, to learn a student network on the unlabeled and unpaired event data (target modality) via knowledge distillation (KD) from a teacher network trained with large labeled image data (source modality). To enable KD across the unpaired modalities, we first propose a bidirectional modality reconstruction (BMR) module to bridge both modalities and simultaneously exploit them to distill knowledge via the crafted pairs, causing no extra computation in the test time. The BMR is improved by the end-task and KD losses in an end-to-end manner. Second, we leverage the structural similarities of both modalities and adapt the knowledge by matching their distributions. Moreover, as most prior feature KD methods are uni-modality and less applicable to our problem, we propose an affinity graph KD and other losses to boost the distillation. Our extensive experiments on semantic segmentation and object recognition demonstrate that EvDistill achieves significantly better results than the prior works and KD with only events and APS frames.
****
LoFTR: Detector-Free Local Feature Matching With Transformers
Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8922-8931
We present a novel method for local image feature matching. Instead of performing image feature detection, description, and matching sequentially, we propose to first establish pixel-wise dense matches at a coarse level and later refine the good matches at a fine level. In contrast to dense methods that use a cost volume to search correspondences, we use self and cross attention layers in Transformer to obtain feature descriptors that are conditioned on both images. The global receptive field provided by Transformer enables our method to produce dense matches in low-texture areas, where feature detectors usually struggle to produce

repeatable interest points. The experiments on indoor and outdoor datasets show that LoFTR outperforms state-of-the-art methods by a large margin. LoFTR also ranks first on two public benchmarks of visual localization among the published methods. Code is available at our project page: https://zju3dv.github.io/loftr/.
****
Combinatorial Learning of Graph Edit Distance via Dynamic Embedding
Runzhong Wang, Tianqi Zhang, Tianshu Yu, Junchi Yan, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5241-5250
Graph Edit Distance (GED) is a popular similarity measurement for pairwise graphs and it also refers to the recovery of the edit path from the source graph to the target graph. Traditional A* algorithm suffers scalability issues due to its exhaustive nature, whose search heuristics heavily rely on human prior knowledge. This paper presents a hybrid approach by combing the interpretability of traditional search-based techniques for producing the edit path, as well as the efficiency and adaptivity of deep embedding models to achieve a cost-effective GED solver. Inspired by dynamic programming, node-level embedding is designated in a dynamic reuse fashion and suboptimal branches are encouraged to be pruned. To this end, our method can be readily integrated into A* procedure in a dynamic fashion, as well as significantly reduce the computational burden with a learned heuristic. Experimental results on different graph datasets show that our approach can remarkably ease the search process of A* without sacrificing much accuracy. To our best knowledge, this work is also the first deep learning-based GED method for recovering the edit path.
****
Radar-Camera Pixel Depth Association for Depth Completion
Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, Praveen Narayanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12507-12516
While radar and video data can be readily fused at the detection level, fusing them at the pixel level is potentially more beneficial. This is also more challenging in part due to the sparsity of radar, but also because automotive radar beams are much wider than a typical pixel combined with a large baseline between camera and radar, which results in poor association between radar pixels and color pixel. A consequence is that depth completion methods designed for LiDAR and video fare poorly for radar and video. Here we propose a radar-to-pixel association stage which learns a mapping from radar returns to pixels. This mapping also serves to densify radar returns. Using this as a first stage, followed by a more traditional depth completion method, we are able to achieve image-guided depth completion with radar and video. We demonstrate performance superior to camera and radar alone on the nuScenes dataset. Our source code is available at https://github.com/longyunf/rc-pda.
****
Improved Image Matting via Real-Time User Clicks and Uncertainty Estimation
Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15374-15383
Image matting is a fundamental and challenging problem in computer vision and graphics. Most existing matting methods leverage a user-supplied trimap as an auxiliary input to produce good alpha matte. However, obtaining high-quality trimap itself is arduous, thus restricting the application of these methods. Recently, some trimap-free methods have emerged, however, the matting quality is still far behind the trimap-based methods. The main reason is that, without the trimap guidance in some cases, the target network is ambiguous about which is the foreground target. In fact, choosing the foreground is a subjective procedure and depends on the user's intention. To this end, this paper proposes an improved deep image matting framework which is trimap-free and only needs several user click interactions to eliminate the ambiguity. Moreover, we introduce a new uncertainty estimation module that can predict which parts need polishing and a following local refinement module. Based on the computation budget, users can choose how many

local parts to improve with the uncertainty guidance. Quantitative and qualitative results show that our method performs better than existing trimap-free methods and comparably to state-of-the-art trimap-based methods with minimal user effort.
****
Revisiting Superpixels for Active Learning in Semantic Segmentation With Realistic Annotation Costs

Lile Cai, Xun Xu, Jun Hao Liew, Chuan Sheng Foo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10988-10997

State-of-the-art methods for semantic segmentation are based on deep neural networks that are known to be data-hungry. Region-based active learning has shown to be a promising method for reducing data annotation costs. A key design choice for region-based AL is whether to use regularly-shaped regions (e.g., rectangles) or irregularly-shaped region (e.g., superpixels). In this work, we address this question under realistic, click-based measurement of annotation costs. In particular, we revisit the use of superpixels and demonstrate that the inappropriate choice of cost measure (e.g., the percentage of labeled pixels), may cause the effectiveness of the superpixel-based approach to be under-estimated. We benchmark the superpixel-based approach against the traditional "rectangle+polygon"-based approach with annotation cost measured in clicks, and show that the former outperforms on both Cityscapes and PASCAL VOC. We further propose a class-balanced acquisition function to boost the performance of the superpixel-based approach and demonstrate its effectiveness on the evaluation datasets. Our results strongly argue for the use of superpixel-based AL for semantic segmentation and highlight the importance of using realistic annotation costs in evaluating such methods.
****
IMODAL: Creating Learnable User-Defined Deformation Models

Leander Lacroix, Benjamin Charlier, Alain Trouve, Barbara Gris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12905-12913

A natural way to model the evolution of an object (growth of a leaf for instance) is to estimate a plausible deforming path between two observations. This interpolation process can generate deceiving results when the set of considered deformations is not relevant to the observed data. To overcome this issue, the framework of deformation modules allows to incorporate in the model structured deformation patterns coming from prior knowledge on the data. The goal of this article is twofold. First defining new deformation modules incorporating structures coming from the elastic properties of the objects. Second, presenting the IMODAL library allowing to perform registration through structured deformations. This library is modular: adapted priors can be easily defined by the user, several priors can be combined into a global one and various types of data can be considered such as curves, meshes or images.
****
Fast End-to-End Learning on Protein Surfaces

Freyr Sverrisson, Jean Feydy, Bruno E. Correia, Michael M. Bronstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15272-15281

Proteins' biological functions are defined by the geometric and chemical structure of their 3D molecular surfaces. Recent works have shown that geometric deep learning can be used on mesh-based representations of proteins to identify potential functional sites, such as binding targets for potential drugs. Unfortunately though, the use of meshes as the underlying representation for protein structure has multiple drawbacks including the need to pre-compute the input features and mesh connectivities. This becomes a bottleneck for many important tasks in protein science. In this paper, we present a new framework for deep learning on protein structures that addresses these limitations. Among the key advantages of our method are the computation and sampling of the molecular surface on-the-fly from the underlying atomic point cloud and a novel efficient geometric convolutional layer. As a result, we are able to process large collections of proteins in a

n end-to-end fashion, taking as the sole input the raw 3D coordinates and chemic
al types of their atoms, eliminating the need for any hand-crafted pre-computed
features. To showcase the performance of our approach, we test it on two tasks i
n the field of protein structural bioinformatics: the identification of interact
ion sites and the prediction of protein-protein interactions. On both tasks, we
achieve state-of-the-art performance with much faster run times and fewer parame
ters than previous models. These results will considerably ease the deployment o
f deep learning methods in protein science and open the door for end-to-end diff
erentiable approaches in protein modeling tasks such as function prediction and
design.
****
Found a Reason for me? Weakly-supervised Grounded Visual Question Answering usin
g Capsules
Aisha Urooj, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, Mubarak Shah; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2021, pp. 8465-8474
The problem of grounding VQA tasks has seen an increased attention in the resear
ch community recently, with most attempts usually focusing on solving this task
by using pretrained object detectors. However, pre-trained object detectors requ
ire bounding box annotations for detecting relevant objects in the vocabulary, w
hich may not always be feasible for real-life large-scale applications. In this
paper, we focus on a more relaxed setting: the grounding of relevant visual enti
ties in a weakly supervised manner by training on the VQA task alone. To address
 this problem, we propose a visual capsule module with a query-based selection m
echanism of capsule features, that allows the model to focus on relevant regions
 based on the textual cues about visual information in the question. We show tha
t integrating the proposed capsule module in existing VQA systems significantly
improves their performance on the weakly supervised grounding task. Overall, we
demonstrate the effectiveness of our approach on two state-of-the-art VQA system
s, stacked NMN and MAC, on the CLEVR-Answers benchmark, our new evaluation set b
ased on CLEVR scenes with ground truth bounding boxes for objects that are relev
ant for the correct answer, as well as on GQA, a real world VQA dataset with com
positional questions. We show that the systems with the proposed capsule module
consistently outperform the respective baseline systems in terms of answer groun
ding, while achieving comparable performance on VQA task.
****
Person Re-Identification Using Heterogeneous Local Graph Attention Networks
Zhong Zhang, Haijia Zhang, Shuang Liu; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12136-12145
Recently, some methods have focused on learning local relation among parts of pe
destrian images for person re-identification (Re-ID), as it offers powerful repr
esentation capabilities. However, they only provide the intra-local relation amo
ng parts within single pedestrian image and ignore the inter-local relation amon
g parts from different images, which results in incomplete local relation inform
ation. In this paper, we propose a novel deep graph model named Heterogeneous Lo
cal Graph Attention Networks (HLGAT) to model the inter-local relation and the i
ntra-local relation in the completed local graph, simultaneously. Specifically,
we first construct the completed local graph using local features, and we resort
 to the attention mechanism to aggregate the local features in the learning proc
ess of inter-local relation and intra-local relation so as to emphasize the impo
rtance of different local features. As for the inter-local relation, we propose
the attention regularization loss to constrain the attention weights based on th
e identities of local features in order to describe the inter-local relation acc
urately. As for the intra-local relation, we propose to inject the contextual in
formation into the attention weights to consider structure information. Extensiv
e experiments on Market-1501, CUHK03, DukeMTMC-reID and MSMT17 demonstrate that
the proposed HLGAT outperforms the state-of-the-art methods.
****
Recurrent Multi-View Alignment Network for Unsupervised Surface Registration
Wanquan Feng, Juyong Zhang, Hongrui Cai, Haofei Xu, Junhui Hou, Hujun Bao; Proce

edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10297-10307

Learning non-rigid registration in an end-to-end manner is challenging due to the inherent high degrees of freedom and the lack of labeled training data. In this paper, we resolve these two challenges simultaneously. First, we propose to represent the non-rigid transformation with a point-wise combination of several rigid transformations. This representation not only makes the solution space well-constrained but also enables our method to be solved iteratively with a recurrent framework, which greatly reduces the difficulty of learning. Second, we introduce a differentiable loss function that measures the 3D shape similarity on the projected multi-view 2D depth images so that our full framework can be trained end-to-end without ground truth supervision. Extensive experiments on several different datasets demonstrate that our proposed method outperforms the previous state-of-the-art by a large margin.

****

Divide-and-Conquer for Lane-Aware Diverse Trajectory Prediction

Sriram Narayanan, Ramin Moslemi, Francesco Pittaluga, Buyu Liu, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15799-15808

Trajectory prediction is a safety-critical tool for autonomous vehicles to plan and execute actions. Our work addresses two key challenges in trajectory prediction, learning multimodal outputs, and better predictions by imposing constraints using driving knowledge. Recent methods have achieved strong performances using Multi-Choice Learning objectives like winner-takes-all (WTA) or best-of-many. But the impact of those methods in learning diverse hypotheses is under-studied as such objectives highly depend on their initialization for diversity. As our first contribution, we propose a novel Divide-And-Conquer (DAC) approach that acts as a better initialization technique to WTA objective, resulting in diverse outputs without any spurious modes. Our second contribution is a novel trajectory prediction framework called ALAN that uses existing lane centerlines as anchors to provide trajectories constrained to the input lanes. Our framework provides multi-agent trajectory outputs in a forward pass by capturing interactions through hypercolumn descriptors and incorporating scene information in the form of rasterized images and per-agent lane anchors. Experiments on synthetic and real data show that the proposed DAC captures the data distribution better compare to other WTA family of objectives. Further, we show that our ALAN approach provides on par or better performance with SOTA methods evaluated on Nuscenes urban driving benchmark.

****

Probabilistic 3D Human Shape and Pose Estimation From Multiple Unconstrained Images in the Wild

Akash Sengupta, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16094-16104

This paper addresses the problem of 3D human body shape and pose estimation from RGB images. Recent progress in this field has focused on single images, video or multi-view images as inputs. In contrast, we propose a new task: shape and pose estimation from a group of multiple images of a human subject, without constraints on subject pose, camera viewpoint or background conditions between images in the group. Our solution to this task predicts distributions over SMPL body shape and pose parameters conditioned on the input images in the group. We probabilistically combine predicted body shape distributions from each image to obtain a final multi-image shape prediction. We show that the additional body shape information present in multi-image input groups improves 3D human shape estimation metrics compared to single-image inputs on the SSP-3D dataset and a private dataset of tape-measured humans. In addition, predicting distributions over 3D bodies allows us to quantify pose prediction uncertainty, which is useful when faced with challenging input images with significant occlusion. Our method demonstrates meaningful pose uncertainty on the 3DPW dataset and is competitive with the state-of-the-art in terms of pose estimation metrics.

****

Weakly Supervised Instance Segmentation for Videos With Temporal Mask Consistenc
y
Qing Liu, Vignesh Ramanathan, Dhruv Mahajan, Alan Yuille, Zhenheng Yang; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2021, pp. 13968-13978
Weakly supervised instance segmentation reduces the cost of annotations required
 to train models. However, existing approaches which rely only on image-level cl
ass labels predominantly suffer from errors due to (a) partial segmentation of o
bjects and (b) missing object predictions. We show that these issues can be bett
er addressed by training with weakly labeled videos instead of images. In videos
, motion and temporal consistency of predictions across frames provide complemen
tary signals which can help segmentation. We are the first to explore the use of
 these video signals to tackle weakly supervised instance segmentation. We propo
se two ways to leverage this information in our model. First, we adapt inter-pix
el relation network (IRN) to effectively incorporate motion information during t
raining. Second, we introduce a new MaskConsist module, which addresses the prob
lem of missing object instances by transferring stable predictions between neigh
boring frames during training. We demonstrate that both approaches together impr
ove the instance segmentation metric AP50 on video frames of two datasets: Youtu
be-VIS and Cityscapes by 5% and 3% respectively.
****
Exploring Data-Efficient 3D Scene Understanding With Contrastive Scene Contexts
Ji Hou, Benjamin Graham, Matthias Niessner, Saining Xie; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 155
87-15597
The rapid progress in 3D scene understanding has come with growing demand for da
ta; however, collecting and annotating 3D scenes (e.g. point clouds) are notorio
usly hard. For example, the number of scenes (e.g. indoor rooms) that can be acc
essed and scanned might be limited; even given sufficient data, acquiring 3D lab
els (e.g. instance masks) requires intensive human labor. In this paper, we expl
ore data-efficient learning for 3D point cloud. As a first step towards this dir
ection, we propose Contrastive Scene Contexts, a 3D pre-training method that mak
es use of both point-level correspondences and spatial contexts in a scene. Our
method achieves state-of-the-art results on a suite of benchmarks where training
 data or labels are scarce. Our study reveals that exhaustive labelling of 3D po
int clouds might be unnecessary; and remarkably, on ScanNet, even using 0.1% of
point labels, we still achieve 89% (instance segmentation) and 96% (semantic seg
mentation) of the baseline performance that uses full annotations.
****
MetaHTR: Towards Writer-Adaptive Handwritten Text Recognition
Ayan Kumar Bhunia, Shuvozit Ghose, Amandeep Kumar, Pinaki Nath Chowdhury, Aneesh
an Sain, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision
and Pattern Recognition (CVPR), 2021, pp. 15830-15839
Handwritten Text Recognition (HTR) remains a challenging problem to date, largel
y due to the varying writing styles that exist amongst us. Prior works however g
enerally operate with the assumption that there is a limited number of styles, m
ost of which have already been captured by existing datasets. In this paper, we
take a completely different perspective -- we work on the assumption that there
is always a new style that is drastically different, and that we will only have
very limited data during testing to perform adaptation. This results in creates
a commercially viable solution -- being exposed to the new style, the model has
the best shot at adaptation, and the few-sample nature makes it practical to imp
lement. We achieve this via a novel meta-learning framework which exploits addit
ional new-writer data via a support set, and outputs a writer-adapted model via
single gradient step update, all during inference. We discover and leverage on t
he important insight that there exists few key characters per writer that exhibi
t relatively larger style discrepancies. For that, we additionally propose to me
ta-learn instance specific weights for a character-wise cross-entropy loss, whic
h is specifically designed to work with the sequential nature of text data. Our
writer-adaptive MetaHTR framework can be easily implemented on the top of most s

tate-of-the-art HTR models. Experiments show an average performance gain of 5-7% can be obtained by observing very few new style data.
****
Learning To Reconstruct High Speed and High Dynamic Range Videos From Events
Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, Ying Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2024-2033
Event cameras are novel sensors that capture the dynamics of a scene asynchronously. Such cameras record event streams with much shorter response latency than images captured by conventional cameras, and are also highly sensitive to intensity change, which is brought by the triggering mechanism of events. On the basis of these two features, previous works attempt to reconstruct high speed and high dynamic range (HDR) videos from events. However, these works either suffer from unrealistic artifacts, or cannot provide sufficiently high frame rate. In this paper, we present a convolutional recurrent neural network which takes a sequence of neighboring events to reconstruct high speed HDR videos, and temporal consistency is well considered to facilitate the training process. In addition, we setup a prototype optical system to collect a real-world dataset with paired high speed HDR videos and event streams, which will be made publicly accessible for future researches in this field. Experimental results on both simulated and real scenes verify that our method can generate high speed HDR videos with high quality, and outperform the state-of-the-art reconstruction methods.
****
PSRR-MaxpoolNMS: Pyramid Shifted MaxpoolNMS With Relationship Recovery
Tianyi Zhang, Jie Lin, Peng Hu, Bin Zhao, Mohamed M. Sabry Aly; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15840-15848
Non-maximum Suppression (NMS) is an essential post-processing step in modern convolutional neural networks for object detection. Unlike convolutions which are inherently parallel, the de-facto standard for NMS, namely GreedyNMS, cannot be easily parallelized and thus could be the performance bottleneck in convolutional object detection pipelines. MaxpoolNMS is introduced as a parallelizable alternative to GreedyNMS, which in turn enables faster speed than GreedyNMS at comparable accuracy. However, MaxpoolNMS is only capable of replacing the GreedyNMS at the first stage of two-stage detectors like Faster-RCNN. There is a significant drop in accuracy when applying MaxpoolNMS at the final detection stage, due to the fact that MaxpoolNMS fails to approximate GreedyNMS precisely in terms of bounding box selection. In this paper, we propose a general, parallelizable and configurable approach PSRR-MaxpoolNMS, to completely replace GreedyNMS at all stages in all detectors. By introducing a simple Relationship Recovery module and a Pyramid Shifted MaxpoolNMS module, our PSRR-MaxpoolNMS is able to approximate GreedyNMS more precisely than MaxpoolNMS. Comprehensive experiments show that our approach outperforms MaxpoolNMS by a large margin, and it is proven faster than GreedyNMS with comparable accuracy. For the first time, PSRR-MaxpoolNMS provides a fully parallelizable solution for customized hardware design, which can be reused for accelerating NMS everywhere.
****
Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset
Zhimeng Zhang, Lincheng Li, Yu Ding, Changjie Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3661-3670
One-shot talking face generation should synthesize high visual quality facial videos with reasonable animations of expression and head pose, and just utilize arbitrary driving audio and arbitrary single face image as the source. Current works fail to generate over 256 x 256 resolution realistic-looking videos due to the lack of an appropriate high-resolution audio-visual dataset, and the limitation of the sparse facial landmarks in providing poor expression details. To synthesize high-definition videos, we build a large in-the-wild high-resolution audio-visual dataset and propose a novel flow-guided talking face generation framework. The new dataset is collected from youtube and consists of about 16 hours 720P

or 1080P videos. We leverage the facial 3D morphable model (3DMM) to split the f ramework into two cascaded modules instead of learning a direct mapping from aud io to video. In the first module, we propose a novel animation generator to prod uce the movements of mouth, eyebrow and head pose simultaneously. In the second module, we transform animation into dense flow to provide more expression detail s and carefully design a novel flow-guided video generator to synthesize videos. Our method is able to produce high-definition videos and outperforms state-of-t he-art works in objective and subjective comparisons.
****
VIGOR: Cross-View Image Geo-Localization Beyond One-to-One Retrieval
Sijie Zhu, Taojiannan Yang, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3640-3649
Cross-view image geo-localization aims to determine the locations of street-view query images by matching with GPS-tagged reference images from aerial view. Rec ent works have achieved surprisingly high retrieval accuracy on city-scale datas ets. However, these results rely on the assumption that there exists a reference image exactly centered at the location of any query image, which is not applica ble for practical scenarios. In this paper, we redefine this problem with a more realistic assumption that the query image can be arbitrary in the area of inter est and the reference images are captured before the queries emerge. This assump tion breaks the one-to-one retrieval setting of existing datasets as the queries and reference images are not perfectly aligned pairs, and there may be multiple reference images covering one query location. To bridge the gap between this re alistic setting and existing datasets, we propose a new large-scale benchmark -- VIGOR-- for cross-View Image Geo-localization beyond One-to-one Retrieval. We be nchmark existing state-of-the-art methods and propose a novel end-to-end framewo rk to localize the query in a coarse-to-fine manner. Apart from the image-level retrieval accuracy, we also evaluate the localization accuracy in terms of the a ctual distance (meters) using the raw GPS data. Extensive experiments are conduc ted under different application scenarios to validate the effectiveness of the p roposed method. The results indicate that cross-view geo-localization in this re alistic setting is still challenging, fostering new research in this direction. Our dataset and code will be publicly available.
****
D-NeRF: Neural Radiance Fields for Dynamic Scenes
Albert Pumarola, Enric Corona, Gerard Pons-Moll, Francesc Moreno-Noguer; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10318-10327
Neural rendering techniques combining machine learning with geometric reasoning have arisen as one of the most promising approaches for synthesizing novel views of a scene from a sparse set of images. Among these, stands out the Neural radi ance fields (NeRF), which trains a deep network to map 5D input coordinates (rep resenting spatial location and viewing direction) into a volume density and view -dependent emitted radiance. However, despite achieving an unprecedented level o f photorealism on the generated images, NeRF is only applicable to static scenes , where the same spatial location can be queried from different images. In this paper we introduce D-NeRF, a method that extends neural radiance fields to a dyn amic domain, allowing to reconstruct and render novel images of objects under ri gid and non-rigid motions. For this purpose we consider time as an additional in put to the system, and split the learning process in two main stages: one that e ncodes the scene into a canonical space and another that maps this canonical rep resentation into the deformed scene at a particular time. Both mappings are lear ned using fully-connected networks. Once the networks are trained, D-NeRF can re nder novel images, controlling both the camera view and the time variable, and t hus, the object movement. We demonstrate the effectiveness of our approach on sc enes with objects under rigid, articulated and non-rigid motions.
****
Towards Unified Surgical Skill Assessment
Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, Ziyu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn

ition (CVPR), 2021, pp. 9522-9531
Surgical skills have a great influence on surgical safety and patients' well-being. Traditional assessment of surgical skills involves strenuous manual efforts, which lacks efficiency and repeatability. Therefore, we attempt to automatically predict how well the surgery is performed using the surgical video. In this paper, a unified multi-path framework for automatic surgical skill assessment is proposed, which takes care of multiple composing aspects of surgical skills, including surgical tool usage, intraoperative event pattern, and other skill proxies. The dependency relationships among these different aspects are specially modeled by a path dependency module in the framework. We conduct extensive experiments on the JIGSAWS dataset of simulated surgical tasks, and a new clinical dataset of real laparoscopic surgeries. The proposed framework achieves promising results on both datasets, with the state-of-the-art on the simulated dataset advanced from 0.71 Spearman's correlation to 0.80. It is also shown that combining multiple skill aspects yields better performance than relying on a single aspect.
****
Read and Attend: Temporal Localisation in Sign Language Videos
Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, Andrew Zisserman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16857-16866
The objective of this work is to annotate sign instances across a broad vocabulary in continuous sign language. We train a Transformer model to ingest a continuous signing stream and output a sequence of written tokens on a large-scale collection of signing footage with weakly-aligned subtitles. We show that through this training it acquires the ability to attend to a large vocabulary of sign instances in the input sequence, enabling their localisation. Our contributions are as follows: (1) we demonstrate the ability to leverage large quantities of continuous signing videos with weakly-aligned subtitles to localise signs in continuous sign language; (2) we employ the learned attention to automatically generate hundreds of thousands of annotations for a large sign vocabulary; (3) we collect a set of 37K manually verified sign instances across a vocabulary of 950 sign classes to support our study of sign language recognition; (4) by training on the newly annotated data from our method, we outperform the prior state of the art on the BSL-1K sign language recognition benchmark.
****
ABMDRNet: Adaptive-Weighted Bi-Directional Modality Difference Reduction Network for RGB-T Semantic Segmentation
Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, Jungong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2633-2642
Semantic segmentation models gain robustness against poor lighting conditions by virtue of complementary information from visible (RGB) and thermal images. Despite its importance, most existing RGB-T semantic segmentation models perform primitive fusion strategies, such as concatenation, element-wise summation and weighted summation, to fuse features from different modalities. These strategies, unfortunately, overlook the modality differences due to different imaging mechanisms, so that they suffer from the reduced discriminability of the fused features. To address such an issue, we propose, for the first time, the strategy of bridging-then-fusing, where the innovation lies in a novel Adaptive-weighted Bi-directional Modality Difference Reduction Network (ABMDRNet). Concretely, a Modality Difference Reduction and Fusion (MDRF) subnetwork is designed, which first employs a bi-directional image-to-image translation based method to reduce the modality differences between RGB features and thermal features, and then adaptively selects those discriminative multi-modality features for RGB-T semantic segmentation in a channel-wise weighted fusion way. Furthermore, considering the importance of contextual information in semantic segmentation, a Multi-Scale Spatial Context (MSC) module and a Multi-Scale Channel Context (MCC) module are proposed to exploit the interactions among multi-scale contextual information of cross-modality features together with their long-range dependencies along spatial and channel dimensions, respectively. Comprehensive experiments on MFNet dataset demonstr

ate that our method achieves new state-of-the-art results.
****
Heterogeneous Grid Convolution for Adaptive, Efficient, and Controllable Computation
Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, Ken Sakurada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13946-13955
This paper proposes a novel heterogeneous grid convolution that builds a graph-based image representation by exploiting heterogeneity in the image content, enabling adaptive, efficient, and controllable computations in a convolutional architecture. More concretely, the approach builds a data-adaptive graph structure from a convolutional layer by a differentiable clustering method, pools features to the graph, performs a novel direction-aware graph convolution, and unpool features back to the convolutional layer. By using the developed module, the paper proposes heterogeneous grid convolutional networks, highly efficient yet strong extension of existing architectures. We have evaluated the proposed approach on four image understanding tasks, semantic segmentation, object localization, road extraction, and salient object detection. The proposed method is effective on three of the four tasks. Especially, the method outperforms a strong baseline with more than 90% reduction in floating-point operations for semantic segmentation, and achieves the state-of-the-art result for road extraction. We will share our code, model, and data.
****
Learning a Facial Expression Embedding Disentangled From Identity
Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, Changjie Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6759-6768
The facial expression analysis requires a compact and identity-ignored expression representation. In this paper, we model the expression as the deviation from the identity by a subtraction operation, extracting a continuous and identity-invariant expression embedding. We propose a Deviation Learning Network (DLN) with a pseudo-siamese structure to extract the deviation feature vector. To reduce the optimization difficulty caused by additional fully connection layers, DLN directly provides high-order polynomial to nonlinearly project the high-dimensional feature to a low-dimensional manifold. Taking label noise into account, we add a crowd layer to DLN for robust embedding extraction. Also, to achieve a more compact representation, we use hierarchical annotation for data augmentation. We evaluate our facial expression embedding on the FEC validation set. The quantitative results prove that we achieve the state-of-the-art, both in terms of fine-grained and identity-invariant property. We further conduct extensive experiments to show that our expression embedding is of high quality for emotion recognition, image retrieval, and face manipulation.
****
Robust Bayesian Neural Networks by Spectral Expectation Bound Regularization
Jiaru Zhang, Yang Hua, Zhengui Xue, Tao Song, Chengyu Zheng, Ruhui Ma, Haibing Guan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3815-3824
Bayesian neural networks have been widely used in many applications because of the distinctive probabilistic representation framework. Even though Bayesian neural networks have been found more robust to adversarial attacks compared with vanilla neural networks, their ability to deal with adversarial noises in practice is still limited. In this paper, we propose Spectral Expectation Bound Regularization (SEBR) to enhance the robustness of Bayesian neural networks. Our theoretical analysis reveals that training with SEBR improves the robustness to adversarial noises. We also prove that training with SEBR can reduce the epistemic uncertainty of the model and hence it can make the model more confident with the predictions, which verifies the robustness of the model from another point of view. Experiments on multiple Bayesian neural network structures and different adversarial attacks validate the correctness of the theoretical findings and the effectiveness of the proposed approach.

****
Learning Probabilistic Ordinal Embeddings for Uncertainty-Aware Regression
Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13896-13905
Uncertainty is the only certainty there is. Modeling data uncertainty is essential for regression, especially in unconstrained settings. Traditionally the direct regression formulation is considered and the uncertainty is modeled by modifying the output space to a certain family of probabilistic distributions. On the other hand, classification based regression and ranking based solutions are more popular in practice while the direct regression methods suffer from the limited performance. How to model the uncertainty within the present-day technologies for regression remains an open issue. In this paper, we propose to learn probabilistic ordinal embeddings which represent each data as a multivariate Gaussian distribution rather than a deterministic point in the latent space. An ordinal distribution constraint is proposed to exploit the ordinal nature of regression. Our probabilistic ordinal embeddings can be integrated into popular regression approaches and empower them with the ability of uncertainty estimation. Experimental results show that our approach achieves competitive performance. Code is available at https://github.com/Li-Wanhua/POEs.
****
StyleMix: Separating Content and Style for Enhanced Data Augmentation
Minui Hong, Jinwoo Choi, Gunhee Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14862-14870
In spite of the great success of deep neural networks for many challenging classification tasks, the learned networks are vulnerable to overfitting and adversarial attacks. Recently, mixup based augmentation methods have been actively studied as one practical remedy for these drawbacks. However, these approaches do not distinguish between the content and style features of the image, but simply mix or cut-and-paste the image. We propose StyleMix and StyleCutMix as the first mixup method that separately manipulates the content and style information of input image pairs. By carefully mixing up the content and style of images, we can create more abundant and robust samples, which eventually enhance the generalization of the model training. We also develop an automatic scheme to decide the degree of style mixing according to the pair's class distance, to prevent messy mixed images from too differently styled pairs. Our experiments on CIFAR-100, CIFAR-10, and ImageNet datasets show that StyleMix achieves comparable performance to state of the art mixup methods and learns more robust classifiers to adversarial attacks.
****
Kaleido-BERT: Vision-Language Pre-Training on Fashion Domain
Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12647-12657
We present a new vision-language (VL) pre-training model dubbed Kaleido-BERT, which introduces a novel kaleido strategy for fashion cross-modality representations from transformers. In contrast to random masking strategy of recent VL models, we design alignment guided masking to jointly focus more on image-text semantic relations. To this end, we carry out five novel tasks, i.e., rotation, jigsaw, camouflage, grey-to-color, and blank-to-color for self-supervised VL pre-training at patches of different scale. Kaleido-BERT is conceptually simple and easy to extend to the existing BERT framework, it attains new state-of-the-art results by large margins on four downstream tasks, including text retrieval (R@1: 4.03% absolute improvement), image retrieval (R@1: 7.13% abs imv.), category recognition (ACC: 3.28% abs imv.), and fashion captioning (Bleu4: 1.2 abs imv.). We validate the efficiency of Kaleido-BERT on a wide range of e-commerical websites, demonstrating its broader potential in real-world applications.
****
Co-Grounding Networks With Semantic Attention for Referring Expression Comprehension in Videos

Sijie Song, Xudong Lin, Jiaying Liu, Zongming Guo, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1346-1355
In this paper, we address the problem of referring expression comprehension in videos, which is challenging due to complex expression and scene dynamics. Unlike previous methods which solve the problem in multiple stages (i.e., tracking, proposal-based matching), we tackle the problem from a novel perspective, co-grounding, with an elegant one-stage framework. We enhance the single-frame grounding accuracy by semantic attention learning and improve the cross-frame grounding consistency with co-grounding feature learning. Semantic attention learning explicitly parses referring cues in different attributes to reduce the ambiguity in the complex expression. Co-grounding feature learning boosts visual feature representations by integrating temporal correlation to reduce the ambiguity caused by scene dynamics. Experiment results demonstrate the superiority of our framework on the video grounding datasets VID and OTB in generating accurate and stable results across frames. Our model is also applicable to referring expression comprehension in images, illustrated by the improved performance on the RefCOCO dataset. Our project is available at https://sijiesong.github.io/co-grounding.
****
Binary Graph Neural Networks
Mehdi Bahri, Gaetan Bahl, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9492-9501
Graph Neural Networks (GNNs) have emerged as a powerful and flexible framework for representation learning on irregular data. As they generalize the operations of classical CNNs on grids to arbitrary topologies, GNNs also bring much of the implementation challenges of their Euclidean counterparts. Model size, memory footprint, and energy consumption are common concerns for many real-world applications. Network binarization allocates a single bit to parameters and activations, thus dramatically reducing the memory requirements (up to 32x compared to single-precision floating-point numbers) and maximizing the benefits of fast SIMD instructions on modern hardware for measurable speedups. However, in spite of the large body of work on binarization for classical CNNs, this area remains largely unexplored in geometric deep learning. In this paper, we present and evaluate different strategies for the binarization of graph neural networks. We show that through careful design of the models, and control of the training process, binary graph neural networks can be trained at only a moderate cost in accuracy on challenging benchmarks. In particular, we present the first dynamic graph neural network in Hamming space, able to leverage efficient k-NN search on binary vectors to speed-up the construction of the dynamic graph. We further verify that the binary models offer significant savings on embedded devices. Our code is publicly available on Github.
****
3D CNNs With Adaptive Temporal Feature Resolutions
Mohsen Fayyaz, Emad Bahrami, Ali Diba, Mehdi Noroozi, Ehsan Adeli, Luc Van Gool, Jurgen Gall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4731-4740
While state-of-the-art 3D Convolutional Neural Networks (CNN) achieve very good results on action recognition datasets, they are computationally very expensive and require many GFLOPs. While the GFLOPs of a 3D CNN can be decreased by reducing the temporal feature resolution within the network, there is no setting that is optimal for all input clips. In this work, we therefore introduce a differentiable Similarity Guided Sampling (SGS) module, which can be plugged into any existing 3D CNN architecture. SGS empowers 3D CNNs by learning the similarity of temporal features and grouping similar features together. As a result, the temporal feature resolution is not anymore static but it varies for each input video clip. By integrating SGS as an additional layer within current 3D CNNs, we can convert them into much more efficient 3D CNNs with adaptive temporal feature resolutions (ATFR). Our evaluations show that the proposed module improves the state-of-the-art by reducing the computational cost (GFLOPs) by half while preserving or even improving the accuracy. We evaluate our module by adding it to multiple s

tate-of-the-art 3D CNNs on various datasets such as Kinetics-600, Kinetics-400, mini-Kinetics, Something-Something V2, UCF101, and HMDB51.
****
Space-Time Neural Irradiance Fields for Free-Viewpoint Video
Wenqi Xian, Jia-Bin Huang, Johannes Kopf, Changil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9421-9431
We present a method that learns a spatiotemporal neural irradiance field for dynamic scenes from a single video. Our learned representation enables free-viewpoint rendering of the input video. Our method builds upon recent advances in implicit representations. Learning a spatiotemporal irradiance field from a single video poses significant challenges because the video contains only one observation of the scene at any point in time. The 3D geometry of a scene can be legitimately represented in numerous ways since varying geometry (motion) can be explained with varying appearance and vice versa. We address this ambiguity by constraining the time-varying geometry of our dynamic scene representation using the scene depth estimated from video depth estimation methods, aggregating contents from individual frames into a single global representation. We provide an extensive quantitative evaluation and demonstrate compelling free-viewpoint rendering results.
****
AutoDO: Robust AutoAugment for Biased Data With Label Noise via Scalable Probabilistic Implicit Differentiation
Denis Gudovskiy, Luca Rigazio, Shun Ishizaka, Kazuki Kozuka, Sotaro Tsukizawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16601-16610
AutoAugment has sparked an interest in automated augmentation methods for deep learning models. These methods estimate image transformation policies for train data that improve generalization to test data. While recent papers evolved in the direction of decreasing policy search complexity, we show that those methods are not robust when applied to biased and noisy data. To overcome these limitations, we reformulate AutoAugment as a generalized automated dataset optimization (AutoDO) task that minimizes the distribution shift between test data and distorted train dataset. In our AutoDO model, we explicitly estimate a set of per-point hyperparameters to flexibly change distribution of train data. In particular, we include hyperparameters for augmentation, loss weights, and soft-labels that are jointly estimated using implicit differentiation. We develop a theoretical probabilistic interpretation of this framework using Fisher information and show that its complexity scales linearly with the dataset size. Our experiments on SVHN, CIFAR-10/100, and ImageNet classification show up to 9.3% improvement for biased datasets with label noise compared to prior methods and, importantly, up to 36.6% gain for underrepresented SVHN classes.
****
Multiple Instance Active Learning for Object Detection
Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5330-5339
Despite the substantial progress of active learning for image recognition, there still lacks an instance-level active learning method specified for object detection. In this paper, we propose Multiple Instance Active Object Detection (MI-AOD), to select the most informative images for detector training by observing instance-level uncertainty. MI-AOD defines an instance uncertainty learning module, which leverages the discrepancy of two adversarial instance classifiers trained on the labeled set to predict instance uncertainty of the unlabeled set. MI-AOD treats unlabeled images as instance bags and feature anchors in images as instances, and estimates the image uncertainty by re-weighting instances in a multiple instance learning (MIL) fashion. Iterative instance uncertainty learning and re-weighting facilitate suppressing noisy instances, toward bridging the gap between instance uncertainty and image-level uncertainty. Experiments validate that MI-AOD sets a solid baseline for instance-level active learning. On commonly use

d object detection datasets, MI-AOD outperforms state-of-the-art methods with si gnificant margins, particularly when the labeled sets are small.
****
Forecasting Irreversible Disease via Progression Learning
Botong Wu, Sijie Ren, Jing Li, Xinwei Sun, Shi-Ming Li, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 021, pp. 8117-8125
Forecasting Parapapillary atrophy (PPA), i.e., a symptom related to most irrever sible eye diseases, provides an alarm for implementing an intervention to slow d own the disease progression at early stage. A key question for this forecast is: how to fully utilize the historical data (e.g., retinal image) up to the curren t stage for future disease prediction? In this paper, we provide an answer with a novel framework, namely Disease Forecast via Progression Learning (DFPL), whic h exploits the irreversibility prior (i.e., cannot be reversed once diagnosed). Specifically, based on this prior, we decompose two factors that contribute to t he prediction of the future disease: i) the current disease label given the data (retinal image, clinical attributes) at present and ii) the future disease labe l given the progression of the retinal images that from the current to the futur e. To model these two factors, we introduce the current and progression predicto rs in DFPL, respectively. In order to account for the degree of progression of t he disease, we propose a temporal generative model to accurately generate the fu ture image and compare it with the current one to get a residual image. The gene rative model is implemented by a recurrent neural network, in order to exploit t he dependency of the historical data. To verify our approach, we apply it to a P PA in-house dataset and it yields a significant improvement (e.g., 4.48% of accu racy; 3.45% of AUC) over others. Besides, our generative model can accurately lo calize the disease-related regions.
****
Understanding the Robustness of Skeleton-Based Action Recognition Under Adversar ial Attack
He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yong-Liang Yang, Kun Zhou, David Hogg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2021, pp. 14656-14665
Action recognition has been heavily employed in many applications such as autono mous vehicles, surveillance, etc, where its robustness is a primary concern. In this paper, we examine the robustness of state-of-the-art action recognizers aga inst adversarial attack, which has been rarely investigated so far. To this end, we propose a new method to attack action recognizers which rely on the 3D skele tal motion. Our method involves an innovative perceptual loss which ensures the imperceptibility of the attack. Empirical studies demonstrate that our method is effective in both white-box and black-box scenarios. Its generalizability is ev idenced on a variety of action recognizers and datasets. Its versatility is show n in different attacking strategies. Its deceitfulness is proven in extensive pe rceptual studies. Our method shows that adversarial attack on 3D skeletal motion s, one type of time-series data, is significantly different from traditional adv ersarial attack problems. Its success raises serious concern on the robustness o f action recognizers and provides insights on potential improvements.
****
Learning Invariant Representations and Risks for Semi-Supervised Domain Adaptati on
Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, Han Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1104-1113
The success of supervised learning crucially hinges on the assumption that train ing data matches test data, which rarely holds in practice due to potential dist ribution shift. In light of this, most existing methods for unsupervised domain adaptation focus on achieving domain-invariant representations and small source domain error. However, recent works have shown that this is not sufficient to gu arantee good generalization on target domain and in fact is provably detrimental under label distribution shift. Furthermore, in many real-world applications it

is often feasible to obtain a small amount of labeled data from the target domain and use them to facilitate model training with source data. Inspired by the above observations, in this paper we propose the first method that aims to simultaneously learn invariant representations and risks under the setting of semi-supervised domain adaptation (Semi-DA). To start with, we first give a finite sample bound for both classification and regression problems under Semi-DA. The bound suggests a principled way for target generalization by aligning both the marginal and conditional distributions across domains in feature space. Motivated by this, we then introduce our LIRR algorithm for jointly Learning Invariant Representations and Risks. Finally, we conduct extensive experiments on both classification and regression tasks to demonstrate the effectiveness of LIRR. Compared with methods that only learn invariant representations or invariant risks, LIRR achieves significant improvements.
****
Cross-MPI: Cross-Scale Stereo for Image Super-Resolution Using Multiplane Images
Yuemei Zhou, Gaochang Wu, Ying Fu, Kun Li, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14842-14851
Various combinations of cameras enrich computational photography, among which reference-based superresolution (RefSR) plays a critical role in multiscale imaging systems. However, existing RefSR approaches fail to accomplish high-fidelity super-resolution under a large resolution gap, e.g., 8x upscaling, due to the lower consideration of the underlying scene structure. In this paper, we aim to solve the RefSR problem in actual multiscale camera systems inspired by multiplane image (MPI) representation. Specifically, we propose Cross-MPI, an end-to-end RefSR network composed of a novel plane-aware attention-based MPI mechanism, a multiscale guided upsampling module as well as a super-resolution (SR) synthesis and fusion module. Instead of using a direct and exhaustive matching between the cross-scale stereo, the proposed plane-aware attention mechanism fully utilizes the concealed scene structure for efficient attention-based correspondence searching. Further combined with a gentle coarse-to-fine guided upsampling strategy, the proposed Cross-MPI can achieve a robust and accurate detail transmission. Experimental results on both digitally synthesized and optical zoom cross-scale data show that the Cross-MPI framework can achieve superior performance against the existing RefSR methods and is a real fit for actual multiscale camera systems even with large-scale differences.
****
Neural Cellular Automata Manifold
Alejandro Hernandez, Armand Vilalta, Francesc Moreno-Noguer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10020-10028
Very recently, the Neural Cellular Automata (NCA) has been proposed to simulate the morphogenesis process with deep networks. NCA learns to grow an image starting from a fixed single pixel. In this work, we show that the neural network (NN) architecture of the NCA can be encapsulated in a larger NN. This allows us to propose a new model that encodes a manifold of NCA, each of them capable of generating a distinct image. Therefore, we are effectively learning an embedding space of CA, which shows generalization capabilities. We accomplish this by introducing dynamic convolutions inside an Auto-Encoder architecture, for the first time used to join two different sources of information, the encoding and cell's environment information. In biological terms, our approach would play the role of the transcription factors, modulating the mapping of genes into specific proteins that drive cellular differentiation, which occurs right before the morphogenesis. We thoroughly evaluate our approach in a dataset of synthetic emojis and also in real images of CIFAR-10. Our model introduces a general-purpose network, which can be used in a broad range of problems beyond image generation.
****
Few-Shot Transformation of Common Actions Into Time and Space
Pengwan Yang, Pascal Mettes, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16031-16040

This paper introduces the task of few-shot common action localization in time and space. Given a few trimmed support videos containing the same but unknown action, we strive for spatio-temporal localization of that action in a long untrimmed query video. We do not require any class labels, interval bounds, or bounding boxes. To address this challenging task, we introduce a novel few-shot transformer architecture with a dedicated encoder-decoder structure optimized for joint commonality learning and localization prediction, without the need for proposals. Experiments on our reorganizations of the AVA and UCF101-24 datasets show the effectiveness of our approach for few-shot common action localization, even when the support videos are noisy. Although we are not specifically designed for common localization in time only, we also compare favorably against the few-shot and one-shot state-of-the-art in this setting. Lastly, we demonstrate that the few-shot transformer is easily extended to common action localization per pixel.
****
MultiLink: Multi-Class Structure Recovery via Agglomerative Clustering and Model Selection
Luca Magri, Filippo Leveni, Giacomo Boracchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1853-1862
We address the problem of recovering multiple structures of different classes in a dataset contaminated by noise and outliers. In particular, we consider geometric structures defined by a mixture of underlying parametric models (e.g. planes and cylinders, homographies and fundamental matrices), and we tackle the robust fitting problem by preference analysis and clustering. We present a new algorithm, termed MultiLink, that simultaneously deals with multiple classes of models. MultiLink wisely combines on-the-fly model fitting and model selection in a novel linkage scheme that determines whether two clusters are to be merged. The resulting method features many practical advantages with respect to methods based on preference analysis, being faster, less sensitive to the inlier threshold, and able to compensate limitations deriving from hypotheses sampling. Experiments on several public datasets demonstrate that MultiLink favorably compares with state of the art alternatives, both in multi-class and single-class problems. Code is publicly made available for download.
****
Meta Pseudo Labels
Hieu Pham, Zihang Dai, Qizhe Xie, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11557-11568
We present Meta Pseudo Labels, a semi-supervised learning method that achieves a new state-of-the-art top-1 accuracy of 90.2% on ImageNet, which is 1.6% better than the existing state-of-the-art. Like Pseudo Labels, Meta Pseudo Labels has a teacher network to generate pseudo labels on unlabeled data to teach a student network. However, unlike Pseudo Labels where the teacher is kept fixed, in Meta Pseudo Labels, the teacher is constantly adapted by the feedback of how well the student performs on the labeled dataset. As a result, the teacher generates better pseudo labels to teach the student.
****
SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction
Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, Gang Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8994-9003
Pedestrian trajectory prediction is a key technology in autopilot, which remains to be very challenging due to complex interactions between pedestrians. However, previous works based on dense undirected interaction suffer from modeling superfluous interactions and neglect of trajectory motion tendency, and thus inevitably result in a considerable deviance from the reality. To cope with these issues, we present a Sparse Graph Convolution Network (SGCN) for pedestrian trajectory prediction. Specifically, the SGCN explicitly models the sparse directed interaction with a sparse directed spatial graph to capture adaptive interaction pedestrians. Meanwhile, we use a sparse directed temporal graph to model the motion tendency, thus to facilitate the prediction based on the observed direction. Finally, parameters of a bi-Gaussian distribution for trajectory prediction are est

imated by fusing the above two sparse graphs. We evaluate our proposed method on the ETH and UCY datasets, and the experimental results show our method outperforms comparative state-of-the-art methods by 9% in Average Displacement Error (ADE) and 13% in Final Displacement Error (FDE). Notably, visualizations indicate that our method can capture adaptive interactions between pedestrians and their effective motion tendencies.
****
Depth Completion Using Plane-Residual Representation
Byeong-Uk Lee, Kyunghyun Lee, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13916-13925
The basic framework of depth completion is to predict a pixel-wise dense depth map using very sparse input data. In this paper, we try to solve this problem in a more effective way, by reformulating the regression-based depth estimation problem into a combination of depth plane classification and residual regression. Our proposed approach is to initially densify sparse depth information by figuring out which plane a pixel should lie among a number of discretized depth planes, and then calculate the final depth value by predicting the distance from the specified plane. This will help the network to lessen the burden of directly regressing the absolute depth information from none, and to effectively obtain more accurate depth prediction result with less computation power and inference time. To do so, we firstly introduce a novel way of interpreting depth information with the closest depth plane label p and a residual value r, as we call it, Plane-Residual (PR) representation. We also propose a depth completion network utilizing PR representation consisting of a shared encoder and two decoders, where one classifies the pixel's depth plane label, while the other one regresses the normalized distance from the classified depth plane. By interpreting depth information in PR representation and using our corresponding depth completion network, we were able to acquire improved depth completion performance with faster computation, compared to previous approaches.
****
Learning an Explicit Weighting Scheme for Adapting Complex HSI Noise
Xiangyu Rui, Xiangyong Cao, Qi Xie, Zongsheng Yue, Qian Zhao, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6739-6748
A general approach for handling hyperspectral image (HSI) denoising issue is to impose weights on different HSI pixels to suppress negative influence brought by noisy elements. Such weighting scheme, however, largely depends on the prior understanding or subjective distribution assumption on HSI noises, making them easily biased to complicated real noises, and hardly generalizable to diverse practical scenarios. Against this issue, this paper proposes a new scheme aiming to capture general weighting principle in a data-driven manner. Specifically, such weighting principle is delivered by an explicit function, called hyperweight-net (HWnet), mapping from an input noisy image to its properly imposed weights. A Bayesian framework, as well as a variational inference algorithm, for inferring HWnet parameters is elaborately designed, expecting to extract the latent weighting rule for general diverse and complicated noisy HSIs. Comprehensive experiments substantiate that the learned HWnet can be not only finely generalized to different noise types from those used in training, but also effectively transferred to other weighted models for the issue. Besides, as a sounder guidance, HWnet can help to more faithfully and robustly achieve deep hyperspectral prior(DHP). Specially, the extracted weights by HWnet are verified to be able to effectively capture complex noise knowledge underlying input HSI, revealing its working insight in experiments.
****
Neural Parts: Learning Expressive 3D Shape Abstractions With Invertible Neural Networks
Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3204-3215
Impressive progress in 3D shape extraction led to representations that can captu

re object geometries with high fidelity. In parallel, primitive-based methods seek to represent objects as semantically consistent part arrangements. However, due to the simplicity of existing primitive representations, these methods fail to accurately reconstruct 3D shapes using a small number of primitives/parts. We address the trade-off between reconstruction quality and number of parts with Neural Parts, a novel 3D primitive representation that defines primitives using an Invertible Neural Network (INN) which implements homeomorphic mappings between a sphere and the target object. The INN allows us to compute the inverse mapping of the homeomorphism, which in turn, enables the efficient computation of both the implicit surface function of a primitive and its mesh, without any additional post-processing. Our model learns to parse 3D objects into semantically consistent part arrangements without any part-level supervision.Evaluations on ShapeNet, D-FAUST and FreiHAND demonstrate that our primitives can capture complex geometries and thus simultaneously achieve geometrically accurate as well as interpretable reconstructions using an order of magnitude fewer primitives than state-of-the-art shape abstraction methods.
****
PV-RAFT: Point-Voxel Correlation Fields for Scene Flow Estimation of Point Clouds
Yi Wei, Ziyi Wang, Yongming Rao, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6954-6963
In this paper, we propose a Point-Voxel Recurrent All-Pairs Field Transforms (PV-RAFT) method to estimate scene flow from point clouds. Since point clouds are irregular and unordered, it is challenging to efficiently extract features from all-pairs fields in the 3D space, where all-pairs correlations play important roles in scene flow estimation. To tackle this problem, we present point-voxel correlation fields, which capture both local and long-range dependencies of point pairs. To capture point-based correlations, we adopt the K-Nearest Neighbors search that preserves fine-grained information in the local region. By voxelizing point clouds in a multi-scale manner, we construct pyramid correlation voxels to model long-range correspondences. Integrating these two types of correlations, our PV-RAFT makes use of all-pairs relations to handle both small and large displacements. We evaluate the proposed method on the FlyingThings3D and KITTI Scene Flow 2015 datasets. Experimental results show that PV-RAFT outperforms state-of-the-art methods by remarkable margins.
****
Improving the Efficiency and Robustness of Deepfakes Detection Through Precise Geometric Features
Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, Weijia Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3609-3618
Deepfakes is a branch of malicious techniques that transplant a target face to the original one in videos, resulting in serious problems such as infringement of copyright, confusion of information, or even public panic. Previous efforts for Deepfakes videos detection mainly focused on appearance features, which have a risk of being bypassed by sophisticated manipulation, also resulting in high model complexity and sensitiveness to noise. Besides, how to mine the temporal features of manipulated videos and exploit them is still an open question. We propose an efficient and robust framework named LRNet for detecting Deepfakes videos through temporal modeling on precise geometric features. A novel calibration module is devised to enhance the precision of geometric features, making it more discriminative, and a two-stream Recurrent Neural Network (RNN) is constructed for sufficient exploitation of temporal features. Compared to previous methods, our proposed method is lighter-weighted and easier to train. Moreover, our method has shown robustness in detecting highly compressed or noise corrupted videos. Our model achieved 0.999 AUC on FaceForensics++ dataset. Meanwhile, it has a graceful decline in performance (-0.042 AUC) when faced with highly compressed videos.
****
Sketch2Model: View-Aware 3D Modeling From Single Free-Hand Sketches

Song-Hai Zhang, Yuan-Chen Guo, Qing-Wen Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6012-6021

We investigate the problem of generating 3D meshes from single free-hand sketches, aiming at fast 3D modeling for novice users. It can be regarded as a single-view reconstruction problem, but with unique challenges, brought by the variation and conciseness of sketches. Ambiguities in poorly-drawn sketches could make it hard to determine how the sketched object is posed. In this paper, we address the importance of viewpoint specification for overcoming such ambiguities, and propose a novel view-aware generation approach. By explicitly conditioning the generation process on a given viewpoint, our method can generate plausible shapes automatically with predicted viewpoints, or with specified viewpoints to help users better express their intentions. Extensive evaluations on various datasets demonstrate the effectiveness of our view-aware design in solving sketch ambiguities and improving reconstruction quality.

****

CASTing Your Model: Learning To Localize Improves Self-Supervised Representations

Ramprasaath R. Selvaraju, Karan Desai, Justin Johnson, Nikhil Naik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11058-11067

Recent advances in self-supervised learning (SSL) have largely closed the gap with supervised ImageNet pretraining. Despite their success these methods have been primarily applied to unlabeled ImageNet images, and show marginal gains when trained on larger sets of uncurated images. We hypothesize that current SSL methods perform best on iconic images, and struggle on complex scene images with many objects. Analyzing contrastive SSL methods shows that they have poor visual grounding and receive poor supervisory signal when trained on scene images. We propose Contrast Attention-Supervised Tuning (CAST) to overcome these limitations. CAST uses unsupervised saliency maps to intelligently sample crops, and to provide grounding supervision via a Grad-CAM attention loss. Experiments on COCO show that CAST significantly improves the features learned by SSL methods on scene images, and further experiments show that CAST-trained models are more robust to changes in backgrounds. We hope that CAST can improve the ability of SSL methods to learn from complex non-iconic images. Our code is available at https://github.com/salesforce/CAST.

****

Robust Consistent Video Depth Estimation

Johannes Kopf, Xuejian Rong, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1611-1621

We present an algorithm for estimating consistent dense depth maps and camera poses from a monocular video. We integrate a learning-based depth prior, in the form of a convolutional neural network trained for single-image depth estimation, with geometric optimization, to estimate a smooth camera trajectory as well as detailed and stable depth reconstruction. Our algorithm combines two complementary techniques: (1) flexible deformation-splines for low-frequency large-scale alignment and (2) geometry-aware depth filtering for high-frequency alignment of fine depth details. In contrast to prior approaches, our method does not require camera poses as input and achieves robust reconstruction for challenging hand-held cell phone captures that contain a significant amount of noise, shake, motion blur, and rolling shutter deformations. Our method quantitatively outperforms state-of-the-arts on the Sintel benchmark for both depth and pose estimations, and attains favorable qualitative results across diverse wild datasets.

****

LaPred: Lane-Aware Prediction of Multi-Modal Future Trajectories of Dynamic Agents

ByeoungDo Kim, Seong Hyeon Park, Seokhwan Lee, Elbek Khoshimjonov, Dongsuk Kum, Junsoo Kim, Jeong Soo Kim, Jun Won Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14636-14645

In this paper, we address the problem of predicting the future motion of a dynamic agent (called a target agent) given its current and past states as well as th

e information on its environment. It is paramount to develop a prediction model that can exploit the contextual information in both static and dynamic environments surrounding the target agent and generate diverse trajectory samples that are meaningful in a traffic context. We propose a novel prediction model, referred to as the lane-aware prediction (LaPred) network, which uses the instance-level lane entities extracted from a semantic map to predict the multimodal future trajectories. For each lane candidate found in the neighborhood of the target agent, LaPred extracts the joint features relating the lane and the trajectories of the neighboring agents. Then, the features for all lane candidates are fused with the attention weights learned through a self-supervised learning task that identifies the lane candidate likely to be followed by the target agent. Using the instance-level lane information, LaPred can produce the trajectories compliant with the surroundings better than 2D raster image-based methods and generate the diverse future trajectories given multiple lane candidates. The experiments conducted on the public nuScenes dataset and Argoverse dataset demonstrate that the proposed LaPred method significantly outperforms the existing prediction models, achieving state-of-the-art performance in the benchmarks.
****

NeuralRecon: Real-Time Coherent 3D Reconstruction From Monocular Video
Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, Hujun Bao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15598-15607
We present a novel framework named NeuralRecon for real-time 3D scene reconstruction from a monocular video. Unlike previous methods that estimate single-view depth maps separately on each key-frame and fuse them later, we propose to directly reconstruct local surfaces represented as sparse TSDF volumes for each video fragment sequentially by a neural network. A learning-based TSDF fusion module based on gated recurrent units is used to guide the network to fuse features from previous fragments. This design allows the network to capture local smoothness prior and global shape prior of 3D surfaces when sequentially reconstructing the surfaces, resulting in accurate, coherent, and real-time surface reconstruction. The experiments on ScanNet and 7-Scenes datasets show that our system outperforms state-of-the-art methods in terms of both accuracy and speed. To the best of our knowledge, this is the first learning-based system that is able to reconstruct dense coherent 3D geometry in real-time. Code is available at the project page: https://zju3dv.github.io/neuralrecon/.
****

Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation
Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4176-4186
While accurate lip synchronization has been achieved for arbitrary-subject audio-driven talking face generation, the problem of how to efficiently drive the head pose remains. Previous methods rely on pre-estimated structural information such as landmarks and 3D parameters, aiming to generate personalized rhythmic movements. However, the inaccuracy of such estimated information under extreme conditions would lead to degradation problems. In this paper, we propose a clean yet effective framework to generate pose-controllable talking faces. We operate on non-aligned raw face images, using only a single photo as an identity reference. The key is to modularize audio-visual representations by devising an implicit low-dimension pose code. Substantially, both speech content and head pose information lie in a joint non-identity embedding space. While speech content information can be defined by learning the intrinsic synchronization between audio-visual modalities, we identify that a pose code will be complementarily learned in a modulated convolution-based reconstruction framework. Extensive experiments show that our method generates accurately lip-synced talking faces whose poses are controllable by other videos. Moreover, our model has multiple advanced capabilities including extreme view robustness and talking face frontalization.
****

Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion
Ho Kei Cheng, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5559-5568
We present Modular interactive VOS (MiVOS) framework which decouples interaction-to-mask and mask propagation, allowing for higher generalizability and better performance. Trained separately, the interaction module converts user interactions to an object mask, which is then temporally propagated by our propagation module using a novel top-k filtering strategy in reading the space-time memory. To effectively take the user's intent into account, a novel difference-aware module is proposed to learn how to properly fuse the masks before and after each interaction, which are aligned with the target frames by employing the space-time memory. We evaluate our method both qualitatively and quantitatively with different forms of user interactions (e.g., scribbles, clicks) on DAVIS to show that our method outperforms current state-of-the-art algorithms while requiring fewer frame interactions, with the additional advantage in generalizing to different types of user interactions. We contribute a large-scale synthetic VOS dataset with pixel-accurate segmentation of 4.8M frames to accompany our source codes to facilitate future research.
****
A Sliced Wasserstein Loss for Neural Texture Synthesis
Eric Heitz, Kenneth Vanhoey, Thomas Chambon, Laurent Belcour; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9412-9420
We address the problem of computing a textural loss based on the statistics extracted from the feature activations of a convolutional neural network optimized for object recognition (e.g. VGG-19). The underlying mathematical problem is the measure of the distance between two distributions in feature space. The Gram-matrix loss is the ubiquitous approximation for this problem but it is subject to several shortcomings. Our goal is to promote the Sliced Wasserstein Distance as a replacement for it. It is theoretically proven, practical, simple to implement, and achieves results that are visually superior for texture synthesis by optimization or training generative neural networks.
****
Learning Accurate Dense Correspondences and When To Trust Them
Prune Truong, Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5714-5724
Establishing dense correspondences between a pair of images is an important and general problem. However, dense flow estimation is often inaccurate in the case of large displacements or homogeneous regions. For most applications and down-stream tasks, such as pose estimation, image manipulation, or 3D reconstruction, it is crucial to know when and where to trust the estimated matches. In this work, we aim to estimate a dense flow field relating two images, coupled with a robust pixel-wise confidence map indicating the reliability and accuracy of the prediction. We develop a flexible probabilistic approach that jointly learns the flow prediction and its uncertainty. In particular, we parametrize the predictive distribution as a constrained mixture model, ensuring better modelling of both accurate flow predictions and outliers. Moreover, we develop an architecture and training strategy tailored for robust and generalizable uncertainty prediction in the context of self-supervised training. Our approach obtains state-of-the-art results on multiple challenging geometric matching and optical flow datasets. We further validate the usefulness of our probabilistic confidence estimation for the task of pose estimation. Code and models are available at https://github.com/PruneTruong/PDCNet.
****
Learning Better Visual Dialog Agents With Pretrained Visual-Linguistic Representation
Tao Tu, Qing Ping, Govindarajan Thattai, Gokhan Tur, Prem Natarajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2

021, pp. 5622-5631
GuessWhat?! is a visual dialog guessing game which incorporates a Questioner agent that generates a sequence of questions, while an Oracle agent answers the respective questions about a target object in an image. Based on this dialog history between the Questioner and the Oracle, a Guesser agent makes a final guess of the target object. While previous work has focused on dialogue policy optimization and visual-linguistic information fusion, most work learns the vision-linguistic encoding for the three agents solely on the GuessWhat?! dataset without shared and prior knowledge of vision-linguistic representation. To bridge these gaps, this paper proposes new Oracle, Guesser and Questioner models that take advantage of a pretrained vision-linguistic model, VilBert. For Oracle model, we introduce a two-way background/target fusion mechanism to understand both intra and inter-object questions. For Guesser model, we introduce a state-estimator that best utilizes Vilbert's strength in single-turn referring expression comprehension. For the Questioner, we share the state-estimator from pretrained Guesser with Questioner to guide the question generator. Experimental results show that our proposed models outperform state-of-the-art models significantly by 7%, 10%, 12% for Oracle, Guesser and End-to-End Questioner respectively.
****
Restoring Extremely Dark Images in Real Time
Mohit Lamba, Kaushik Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3487-3497
A practical low-light enhancement solution must be computationally fast, memory-efficient, and achieve a visually appealing restoration. Most of the existing methods target restoration quality and thus compromise on speed and memory requirements, raising concerns about their real-world deployability. We propose a new deep learning architecture for extreme low-light single image restoration, which is exceptionally lightweight, remarkably fast, and produces a restoration that is perceptually at par with state-of-the-art computationally intense models. To achieve this, we do most of the processing in the higher scale-spaces, skipping the intermediate-scales wherever possible. Also unique to our model is the potential to process all the scale-spaces concurrently, offering an additional 30% speedup without compromising the restoration quality. Pre-amplification of the dark raw-image is an important step in extreme low-light image enhancement. Most of the existing state-of-the-art methods need GT exposure value to estimate the pre-amplification factor, which is not practically feasible. Thus, we propose an amplifier module that estimates the amplification factor using only the input raw image and can be used "off-the-shelf"" with pre-trained models without any fine-tuning. We show that our model can restore an ultra-high-definition 4K resolution image in just 1sec on a CPU and at 32fps on a GPU and yet maintain a competitive restoration quality. We also show that our proposed model, without any fine-tuning, generalizes well to cameras not seen during training and to subsequent tasks such as object detection.
****
Weakly-Supervised Instance Segmentation via Class-Agnostic Learning With Salient Images
Xinggang Wang, Jiapei Feng, Bin Hu, Qi Ding, Longjin Ran, Xiaoxin Chen, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10225-10235
Humans have a strong class-agnostic object segmentation ability and can outline boundaries of unknown objects precisely, which motivates us to propose a box-supervised class-agnostic object segmentation (BoxCaseg) based solution for weakly-supervised instance segmentation. The BoxCaseg model is jointly trained using box-supervised images and salient images in a multi-task learning manner. The fine-annotated salient images provide class-agnostic and precise object localization guidance for box-supervised images. The object masks predicted by a pretrained BoxCaseg model are refined via a novel merged and dropped strategy as proxy ground truth to train a Mask R-CNN for weakly-supervised instance segmentation. Only using 7991 salient images, the weakly-supervised Mask R-CNN is on par with fully-supervised Mask R-CNN on PASCAL VOC and significantly outperforms previous sta

te-of-the-art box-supervised instance segmentation methods on COCO. The source code, pretrained models and datasets are available at https://github.com/hustvl/BoxCaseg.
****
Spoken Moments: Learning Joint Audio-Visual Representations From Video Descriptions
Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, Aude Oliva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14871-14881
When people observe events, they are able to abstract key information and build concise summaries of what is happening. These summaries include contextual and semantic information describing the important high-level details (what, where, who and how) of the observed event and exclude background information that is deemed unimportant to the observer. With this in mind, the descriptions people generate for videos of different dynamic events can greatly improve our understanding of the key information of interest in each video. These descriptions can be captured in captions that provide expanded attributes for video labeling (e.g. actions/objects/scenes/sentiment/etc.) while allowing us to gain new insight into what people find important or necessary to summarize specific events. Existing caption datasets for video understanding are either small in scale or restricted to a specific domain. To address this, we present the Spoken Moments (S-MiT) dataset of 500k spoken captions each attributed to a unique short video depicting a broad range of different events. We collect our descriptions using audio recordings to ensure that they remain as natural and concise as possible while allowing us to scale the size of a large classification dataset. In order to utilize our proposed dataset, we present a novel Adaptive Mean Margin (AMM) approach to contrastive learning and evaluate our models on video/caption retrieval on multiple datasets. We show that our AMM approach consistently improves our results and that models trained on our Spoken Moments dataset generalize better than those trained on other video-caption datasets.
****
Image Restoration for Under-Display Camera
Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, Timothy Large; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9179-9188
The new trend of full-screen devices encourages us to position a camera behind a screen. Removing the bezel and centralizing the camera under the screen brings larger display-to-body ratio and enhances eye contact in video chat, but also causes image degradation. In this paper, we focus on a newly-defined Under-Display Camera (UDC), as a novel real-world single image restoration problem. First, we take a 4k Transparent OLED (T-OLED) and a phone Pentile OLED (P-OLED) and analyze their optical systems to understand the degradation. Second, we design a Monitor-Camera Imaging System (MCIS) for easier real pair data acquisition, and a model-based data synthesizing pipeline to generate Point Spread Function (PSF) and UDC data only from display pattern and camera measurements. Finally, we resolve the complicated degradation using deconvolution-based pipeline and learning-based methods. Our model demonstrates a real-time high-quality restoration. The presented methods and results reveal the promising research values and directions of UDC.
****
Unbiased Mean Teacher for Cross-Domain Object Detection
Jinhong Deng, Wen Li, Yuhua Chen, Lixin Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4091-4101
Cross-domain object detection is challenging, because object detection model is often vulnerable to data variance, especially to the considerable domain shift between two distinctive domains. In this paper, we propose a new Unbiased Mean Teacher (UMT) model for cross-domain object detection. We reveal that there often exists a considerable model bias for the simple mean teacher (MT) model in cross-domain scenarios, and eliminate the model bias with several simple yet highly effective strategies. In particular, for the teacher model, we propose a cross-do

main distillation for MT to maximally exploit the expertise of the teacher model. Second, for the student model, we also alleviate its bias by augmenting training samples with pixel-level adaptation. Finally, for the teaching process, we employ an out-of-distribution estimation strategy to select samples that most fit the current model to further enhance the cross-domain distillation process. By tackling the model bias issue with these strategies, our UMT model achieves mAPs of 44.1%, 58.1%, 41.7%, and 43.1% on benchmark datasets Clipart1k, Watercolor2k, Foggy Cityscapes, and Cityscapes, respectively, which outperforms the existing state-of-the-art results in notable margins. Our implementation is available at https://github.com/kinredon/umt.
****

How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language
Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, Xavier Giro-i-Nieto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2735-2744
One of the factors that have hindered progress in the areas of sign language recognition, translation, and production is the absence of large annotated datasets. Towards this end, we introduce How2Sign, a multimodal and multiview continuous American Sign Language (ASL) dataset, consisting of a parallel corpus of more than 80 hours of sign language videos and a set of corresponding modalities including speech, English transcripts, and depth. A three-hour subset was further recorded in the Panoptic studio enabling detailed 3D pose estimation. To evaluate the potential of How2Sign for real-world impact, we conduct a study with ASL signers and show that synthesized videos using our dataset can indeed be understood. The study further gives insights on challenges that computer vision should address in order to make progress in this field. Dataset website: http://how2sign.github.io/
****

Indoor Lighting Estimation Using an Event Camera
Zehao Chen, Qian Zheng, Peisong Niu, Huajin Tang, Gang Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14760-14770
Image-based methods for indoor lighting estimation suffer from the problem of intensity-distance ambiguity. This paper introduces a novel setup to help alleviate the ambiguity based on the event camera. We further demonstrate that estimating the distance of a light source becomes a well-posed problem under this setup, based on which an optimization-based method and a learning-based method are proposed. Our experimental results validate that our approaches not only achieve superior performance for indoor lighting estimation (especially for the close light) but also significantly alleviate the intensity-distance ambiguity.
****

Shot Contrastive Self-Supervised Learning for Scene Boundary Detection
Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, Raffay Hamid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9796-9805
Scenes play a crucial role in breaking the storyline of movies and TV episodes into semantically cohesive parts. However, given their complex temporal structure, finding scene boundaries can be a challenging task requiring large amounts of labeled training data. To address this challenge, we present a self-supervised shot contrastive learning approach (ShotCoL) to learn a shot representation that maximizes the similarity between nearby shots compared to randomly selected shots. We show how to apply our learned shot representation for the task of scene boundary detection to offer state-of-the-art performance on the MovieNet dataset while requiring only 25% of the training labels, using 9x fewer model parameters and offering 7x faster runtime. To assess the effectiveness of ShotCoL on novel applications of scene boundary detection, we take on the problem of finding timestamps in movies and TV episodes where video-ads can be inserted while offering a minimally disruptive viewing experience. To this end, we collected a new dataset called AdCuepoints with 3,975 movies and TV episodes, 2.2 million shots and

19,119 minimally disruptive ad cue-point labels. We present a thorough empirical analysis on this dataset demonstrating the effectiveness of ShotCoL for ad cue-points detection.
****
Sewer-ML: A Multi-Label Sewer Defect Classification Dataset and Benchmark
Joakim Bruslund Haurum, Thomas B. Moeslund; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13456-13467
Perhaps surprisingly sewerage infrastructure is one of the most costly infrastructures in modern society. Sewer pipes are manually inspected to determine whether the pipes are defective. However, this process is limited by the number of qualified inspectors and the time it takes to inspect a pipe. Automatization of this process is therefore of high interest. So far, the success of computer vision approaches for sewer defect classification has been limited when compared to the success in other fields mainly due to the lack of public datasets. To this end, in this work we present a large novel and publicly available multi-label classification dataset for image-based sewer defect classification called Sewer-ML. The Sewer-ML dataset consists of 1.3 million images annotated by professional sewer inspectors from three different utility companies across nine years. Together with the dataset, we also present a benchmark algorithm and a novel metric for assessing performance. The benchmark algorithm is a result of evaluating 12 state-of-the-art algorithms, six from the sewer defect classification domain and six from the multi-label classification domain, and combining the best performing algorithms. The novel metric is a class-importance weighted F2 score, F2-CIW, reflecting the economic impact of each class, used together with the normal pipe F1 score, F1-Normal. The benchmark algorithm achieves an F2-CIW score of 55.11% and F1-Normal score of 90.94%, leaving ample room for improvement on the Sewer-ML dataset. The code, models, and dataset are available at the project page http://vap.aau.dk/sewer-ml
****
Joint-DetNAS: Upgrade Your Detector With NAS, Pruning and Dynamic Distillation
Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, Tong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10175-10184
We propose Joint-DetNAS, a unified NAS framework for object detection, which integrates 3 key components: Neural Architecture Search, pruning, and Knowledge Distillation. Instead of naively pipelining these techniques, our Joint-DetNAS optimizes them jointly. The algorithm consists of two core processes: student morphism optimizes the student's architecture and removes the redundant parameters, while dynamic distillation aims to find the optimal matching teacher. For student morphism, weight inheritance strategy is adopted, allowing the student to flexibly update its architecture while fully utilize the predecessor's weights, which considerably accelerates the search; To facilitate dynamic distillation, an elastic teacher pool is trained via integrated progressive shrinking strategy, from which teacher detectors can be sampled without additional cost in subsequent searches. Given a base detector as the input, our algorithm directly outputs the derived student detector with high performance without additional training. Experiments demonstrate that our Joint-DetNAS outperforms the naive pipelining approach by a great margin. Given a classic R101-FPN as the base detector, Joint-DetNAS is able to boost its mAP from 41.4 to 43.9 on MS COCO and reduce the latency by 47%, which is on par with the SOTA EfficientDet while requiring less search cost. We hope our proposed method can provide the community with a new way of jointly optimizing NAS, KD and pruning.
****
Back-Tracing Representative Points for Voting-Based 3D Object Detection in Point Clouds
Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8963-8972
3D object detection in point clouds is a challenging vision task that benefits various applications for understanding the 3D visual world. Lots of recent resear

ch focuses on how to exploit end-to-end trainable Hough voting for generating ob ject proposals. However, the current voting strategy can only receive partial vo tes from the surfaces of potential objects together with severe outlier votes fr om the cluttered backgrounds, which hampers full utilization of the information from the input point clouds. Inspired by the back-tracing strategy in the conven tional Hough voting methods, in this work, we introduce a new 3D object detectio n method, named as Back-tracing Representative Points Network (BRNet), which gen eratively back-traces the representative points from the vote centers and also r evisits complementary seed points around these generated points, so as to better capture the fine local structural features surrounding the potential objects fr om the raw point clouds. Therefore, this bottom-up and then top-down strategy in our BRNet enforces mutual consistency between the predicted vote centers and th e raw surface points and thus achieves more reliable and flexible object localiz ation and class prediction results. Our BRNet is simple but effective, which sig nificantly outperforms the state-of-the-art methods on two large-scale point clo ud datasets, ScanNet V2 (+7.5% in terms of mAP@0.50) and SUN RGB-D (+4.7% in ter ms of mAP@0.50), while it is still lightweight and efficient.
****

High-Resolution Photorealistic Image Translation in Real-Time: A Laplacian Pyram id Translation Network

Jie Liang, Hui Zeng, Lei Zhang; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2021, pp. 9392-9400

Existing image-to-image translation (I2IT) methods are either constrained to low -resolution images or long inference time due to their heavy computational burde n on the convolution of high-resolution feature maps. In this paper, we focus on speeding-up the high-resolution photorealistic I2IT tasks based on closed-form Laplacian pyramid decomposition and reconstruction. Specifically, we reveal that the attribute transformations, such as illumination and color manipulation, rel ate more to the low-frequency component, while the content details can be adapti vely refined on high-frequency components. We consequently propose a Laplacian P yramid Translation Network (LPTN) to simultaneously perform these two tasks, whe re we design a lightweight network for translating the low-frequency component w ith reduced resolution and a progressive masking strategy to efficiently refine the high-frequency ones. Our model avoids most of the heavy computation consumed by processing high-resolution feature maps and faithfully preserves the image d etails. Extensive experimental results on various tasks demonstrate that the pro posed method can translate 4K images in real-time using one normal GPU while ach ieving comparable transformation performance against existing methods. Datasets and codes are available: https://github.com/csjliang/LPTN.
****

End-to-End Video Instance Segmentation With Transformers

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, Huaxia Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2021, pp. 8741-8750

Video instance segmentation (VIS) is the task that requires simultaneously class ifying, segmenting and tracking object instances of interest in video. Recent me thods typically develop sophisticated pipelines to tackle this task. Here, we pr opose a new video instance segmentation framework built upon Transformers, terme d VisTR, which views the VIS task as a direct end-to-end parallel sequence decod ing/prediction problem. Given a video clip consisting of multiple image frames a s input, VisTR outputs the sequence of masks for each instance in the video in o rder directly. At the core is a new, effective instance sequence matching and se gmentation strategy, which supervises and segments instances at the sequence lev el as a whole. VisTR frames the instance segmentation and tracking in the same p erspective of similarity learning, thus considerably simplifying the overall pip eline and is significantly different from existing approaches. Without bells and whistles, VisTR achieves the highest speed among all existing VIS models, and a chieves the best result among methods using single model on the YouTube-VIS data set. For the first time, we demonstrate a much simpler and faster video instance segmentation framework achieving competitive accuracy. We hope that VisTR can m

otivate future research for more video understanding tasks. Code is available at : https://git.io/VisTR
****
VoxelContext-Net: An Octree Based Framework for Point Cloud Compression
Zizheng Que, Guo Lu, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6042-6051
In this paper, we propose a two-stage deep learning framework called VoxelContext-Net for both static and dynamic point cloud compression. Taking advantages of both octree based methods and voxel based schemes, our approach employs the voxel context to compress the octree structured data. Specifically, we first extract the local voxel representation that encodes the spatial neighbouring context information for each node in the constructed octree. Then, in the entropy coding stage, we propose a voxel context based deep entropy model to compress the symbols of non-leaf nodes in a lossless way. Furthermore, for dynamic point cloud compression, we additionally introduce the local voxel representations from the temporal neighbouring point clouds to exploit temporal dependency. More importantly, to alleviate the distortion from the octree construction procedure, we propose a voxel context based 3D coordinate refinement method to produce more accurate reconstructed point cloud at the decoder side, which is applicable to both static and dynamic point cloud compression. The comprehensive experiments on both static and dynamic point cloud benchmark datasets(e.g., ScanNet and Semantic KITTI) clearly demonstrate the effectiveness of our newly proposed method VoxelContext-Net for 3D point cloud geometry compression.
****
A Second-Order Approach to Learning With Instance-Dependent Label Noise
Zhaowei Zhu, Tongliang Liu, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10113-10123
The presence of label noise often misleads the training of deep neural networks. Departing from the recent literature which largely assumes the label noise rate is only determined by the true label class, the errors in human-annotated labels are more likely to be dependent on the difficulty levels of tasks, resulting in settings with instance-dependent label noise. We first provide evidences that the heterogeneous instance-dependent label noise is effectively down-weighting the examples with higher noise rates in a non-uniform way and thus causes imbalances, rendering the strategy of directly applying methods for class-dependent label noise questionable. Built on a recent work peer loss [24], we then propose and study the potentials of a second-order approach that leverages the estimation of several covariance terms defined between the instance-dependent noise rates and the Bayes optimal label. We show that this set of second-order statistics successfully captures the induced imbalances. We further proceed to show that with the help of the estimated second-order statistics, we identify a new loss function whose expected risk of a classifier under instance-dependent label noise is equivalent to a new problem with only class-dependent label noise. This fact allows us to apply existing solutions to handle this better-studied setting. We provide an efficient procedure to estimate these second-order statistics without accessing either ground truth labels or prior knowledge of the noise rates. Experiments on CIFAR10 and CIFAR100 with synthetic instance-dependent label noise and Clothing1M with real-world human label noise verify our approach. Our implementation is available at https://github.com/UCSC-REAL/CAL.
****
SpinNet: Learning a General Surface Descriptor for 3D Point Cloud Registration
Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11753-11762
Extracting robust and general 3D local features is key to downstream tasks such as point cloud registration and reconstruction. Existing learning-based local descriptors are either sensitive to rotation transformations, or rely on classical handcrafted features which are neither general nor representative. In this paper, we introduce a new, yet conceptually simple, neural architecture, termed SpinNet, to extract local features which are rotationally invariant whilst sufficien

tly informative to enable accurate registration. A Spatial Point Transformer is first introduced to map the input local surface into a carefully designed cylind rical space, enabling end-to-end optimization with SO(2) equivariant representat ion. A Neural Feature Extractor which leverages the powerful point-based and 3D cylindrical convolutional neural layers is then utilized to derive a compact and representative descriptor for matching. Extensive experiments on both indoor an d outdoor datasets demonstrate that SpinNet outperforms existing state-of-the-ar t techniques by a large margin. More critically, it has the best generalization ability across unseen scenarios with different sensor modalities.
****
FSDR: Frequency Space Domain Randomization for Domain Generalization
Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6891-6902
Domain generalization aims to learn a generalizable model from a `known' source domain for various `unknown' target domains. It has been studied widely by domai n randomization that transfers source images to different styles in spatial spac e for learning domain-agnostic features. However, most existing randomization me thods use GANs that often lack of controls and even alter semantic structures of images undesirably. Inspired by the idea of JPEG that converts spatial images i nto multiple frequency components (FCs), we propose Frequency Space Domain Rando mization (FSDR) that randomizes images in frequency space by keeping domain-inva riant FCs (DIFs) and randomizing domain-variant FCs (DVFs) only. FSDR has two un ique features: 1) it decomposes images into DIFs and DVFs which allows explicit access and manipulation of them and more controllable randomization; 2) it has m inimal effects on semantic structures of images and domain-invariant features. W e examined domain variance and invariance property of FCs statistically and desi gned a network that can identify and fuse DIFs and DVFs dynamically through iter ative learning. Extensive experiments over multiple domain generalizable segment ation tasks show that FSDR achieves superior segmentation and its performance is even on par with domain adaptation methods that access target data in training.
****
DualAST: Dual Style-Learning Networks for Artistic Style Transfer
Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xi ng, Dongming Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2021, pp. 872-881
Artistic style transfer is an image editing task that aims at repainting everyda y photographs with learned artistic styles. Existing methods learn styles from e ither a single style example or a collection of artworks. Accordingly, the styli zation results are either inferior in visual quality or limited in style control lability. To tackle this problem, we propose a novel Dual Style-Learning Artisti c Style Transfer (DualAST) framework to learn simultaneously both the holistic a rtist-style (from a collection of artworks) and the specific artwork-style (from a single style image): the artist-style sets the tone (i.e., the overall feelin g) for the stylized image, while the artwork-style determines the details of the stylized image, such as color and texture. Moreover, we introduce a Style-Contr ol Block (SCB) to adjust the styles of generated images with a set of learnable style-control factors. We conduct extensive experiments to evaluate the performa nce of the proposed framework, the results of which confirm the superiority of o ur method.
****
Learning a Proposal Classifier for Multiple Object Tracking
Peng Dai, Renliang Weng, Wongun Choi, Changshui Zhang, Zhangping He, Wei Ding; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2443-2452
The recent trend in multiple object tracking (MOT) is heading towards leveraging deep learning to boost the tracking performance. However, it is not trivial to solve the data-association problem in an end-to-end fashion. In this paper, we p ropose a novel proposal-based learnable framework, which models MOT as a proposa l generation, proposal scoring and trajectory inference paradigm on an affinity graph. This framework is similar to the two-stage object detector Faster RCNN, a

nd can solve the MOT problem in a data-driven way. For proposal generation, we propose an iterative graph clustering method to reduce the computational cost while maintaining the quality of the generated proposals. For proposal scoring, we deploy a trainable graph-convolutional-network (GCN) to learn the structural patterns of the generated proposals and rank them according to the estimated quality scores. For trajectory inference, a simple deoverlapping strategy is adopted to generate tracking output while complying with the constraints that no detection can be assigned to more than one track. We experimentally demonstrate that the proposed method achieves a clear performance improvement in both MOTA and IDF1 with respect to previous state-of-the-art on two public benchmarks. Our code is available at https://github.com/daip13/LPC_MOT.git.
****

Multi-Attentional Deepfake Detection
Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2185-2194
Face forgery by deepfake is widely spread over the internet and has raised severe societal concerns. Recently, how to detect such forgery contents has become a hot research topic and many deepfake detection methods have been proposed. Most of them model deepfake detection as a vanilla binary classification problem, i.e, first use a backbone network to extract a global feature and then feed it into a binary classifier (real/fake). But since the difference between the real and fake images in this task is often subtle and local, we argue this vanilla solution is not optimal. In this paper, we instead formulate deepfake detection as a fine-grained classification problem and propose a new multi-attentional deepfake detection network. Specifically, it consists of three key components: 1) multiple spatial attention heads to make the network attend to different local parts; 2) textural feature enhancement block to zoom in the subtle artifacts in shallow features; 3) aggregate the low-level textural feature and high-level semantic features guided by the attention maps. Moreover, to address the learning difficulty of this network, we further introduce a new regional independence loss and an attention guided data augmentation strategy. Through extensive experiments on different datasets, we demonstrate the superiority of our method over the vanilla binary classifier counterparts, and achieve state-of-the-art performance.
****

SOLD2: Self-Supervised Occlusion-Aware Line Description and Detection
Remi Pautrat, Juan-Ting Lin, Viktor Larsson, Martin R. Oswald, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11368-11378
Compared to feature point detection and description, detecting and matching line segments offer additional challenges. Yet, line features represent a promising complement to points for multi-view tasks. Lines are indeed well-defined by the image gradient, frequently appear even in poorly textured areas and offer robust structural cues. We thus hereby introduce the first joint detection and description of line segments in a single deep network. Thanks to a self-supervised training, our method does not require any annotated line labels and can therefore generalize to any dataset. Our detector offers repeatable and accurate localization of line segments in images, departing from the wireframe parsing approach. Leveraging the recent progresses in descriptor learning, our proposed line descriptor is highly discriminative, while remaining robust to viewpoint changes and occlusions. We evaluate our approach against previous line detection and description methods on several multi-view datasets created with homographic warps as well as real-world viewpoint changes. Our full pipeline yields higher repeatability, localization accuracy and matching metrics, and thus represents a first step to bridge the gap with learned feature points methods. Code and trained weights are available at https://github.com/cvg/SOLD2.
****

Shared Cross-Modal Trajectory Prediction for Autonomous Driving
Chiho Choi, Joon Hee Choi, Jiachen Li, Srikanth Malla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 244-2

Predicting future trajectories of traffic agents in highly interactive environme
nts is an essential and challenging problem for the safe operation of autonomous
 driving systems. On the basis of the fact that self-driving vehicles are equipp
ed with various types of sensors (e.g., LiDAR scanner, RGB camera, radar, etc.),
 we propose a Cross-Modal Embedding framework that aims to benefit from the use
of multiple input modalities. At training time, our model learns to embed a set
of complementary features in a shared latent space by jointly optimizing the obj
ective functions across different types of input data. At test time, a single in
put modality (e.g., LiDAR data) is required to generate predictions from the inp
ut perspective (i.e., in the LiDAR space), while taking advantages from the mode
l trained with multiple sensor modalities. An extensive evaluation is conducted
to show the efficacy of the proposed framework using two benchmark driving datas
ets.
****
Cycle4Completion: Unpaired Point Cloud Completion Using Cycle Transformation Wit
h Missing Region Coding
Xin Wen, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, Yu-Shen Liu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2021, pp. 13080-13089
In this paper, we present a novel unpaired point cloud completion network, named
 Cycle4Completion, to infer the complete geometries from a partial 3D object. Pr
evious unpaired completion methods merely focus on the learning of geometric cor
respondence from incomplete shapes to complete shapes, and ignore the learning i
n the reverse direction, which makes them suffer from low completion accuracy du
e to the limited 3D shape understanding ability. To address this problem, we pro
pose two simultaneous cycle transformations between the latent spaces of complet
e shapes and incomplete ones. Specifically, the first cycle transforms shapes fr
om incomplete domain to complete domain, and then projects them back to the inco
mplete domain. This process learns the geometric characteristic of complete shap
es, and maintains the shape consistency between the complete prediction and the
incomplete input. Similarly, the inverse cycle transformation starts from comple
te domain to incomplete domain, and goes back to complete domain to learn the ch
aracteristic of incomplete shapes. We experimentally show that our model with th
e learned bidirectional geometry correspondence outperforms state-of-the-art unp
aired completion methods. Code will be available at https://github.com/diviswen/
Cycle4Completion.
****
CGA-Net: Category Guided Aggregation for Point Cloud Semantic Segmentation
Tao Lu, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2021, pp. 11693-11702
Previous point cloud semantic segmentation networks use the same process to aggr
egate features from neighbors of the same category and different categories. How
ever, the joint area between two objects usually only occupies a small percentag
e in the whole scene. Thus the networks are well-trained for aggregating feature
s from the same category point while not fully trained on aggregating points of
different categories. To address this issue, this paper proposes to utilize diff
erent aggregation strategies between the same category and different categories.
 Specifically, it presents a customized module, termed as Category Guided Aggreg
ation (CGA), where it first identifies whether the neighbors belong to the same
category with the center point or not, and then handles the two types of neighbo
rs with two carefully-designed modules. Our CGA presents a general network modul
e and could be leveraged in any existing semantic segmentation network. Experime
nts on three different backbones demonstrate the effectiveness of our method.
****
PLOP: Learning Without Forgetting for Continual Semantic Segmentation
Arthur Douillard, Yifu Chen, Arnaud Dapogny, Matthieu Cord; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
4040-4050
Deep learning approaches are nowadays ubiquitously used to tackle computer visio

n tasks such as semantic segmentation, requiring large datasets and substantial computational power. Continual learning for semantic segmentation (CSS) is an emerging trend that consists in updating an old model by sequentially adding new classes. However, continual learning methods are usually prone to catastrophic forgetting. This issue is further aggravated in CSS where, at each step, old classes from previous iterations are collapsed into the background. In this paper, we propose Local POD, a multi-scale pooling distillation scheme that preserves long- and short-range spatial relationships at feature level. Furthermore, we design an entropy-based pseudo-labelling of the background w.r.t. classes predicted by the old model to deal with background shift and avoid catastrophic forgetting of the old classes. Our approach, called PLOP, significantly outperforms state-of-the-art methods in existing CSS scenarios, as well as in newly proposed challenging benchmarks.
****
Magic Layouts: Structural Prior for Component Detection in User Interface Designs
Dipu Manandhar, Hailin Jin, John Collomosse; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15809-15818
We present Magic Layouts; a method for parsing screenshots or hand-drawn sketches of user interface (UI) layouts. Our core contribution is to extend existing detectors to exploit a learned structural prior for UI designs, enabling robust detection of UI components; buttons, text boxes and similar. Specifically we learn a prior over mobile UI layouts, encoding common spatial co-occurrence relationships between different UI components. Conditioning region proposals using this prior leads to performance gains on UI layout parsing for both hand-drawn UIs and app screenshots, which we demonstrate within the context an interactive application for rapidly acquiring digital prototypes of user experience (UX) designs.
****
MetaAlign: Coordinating Domain Alignment and Classification for Unsupervised Domain Adaptation
Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16643-16653
For unsupervised domain adaptation (UDA), to alleviate the effect of domain shift, many approaches align the source and target domains in the feature space by adversarial learning or by explicitly aligning their statistics. However, the optimization objective of such domain alignment is generally not coordinated with that of the object classification task itself such that their descent directions for optimization may be inconsistent. This will reduce the effectiveness of domain alignment in improving the performance of UDA. In this paper, we aim to study and alleviate the optimization inconsistency problem between the domain alignment and classification tasks. We address this by proposing an effective meta-optimization based strategy dubbed MetaAlign, where we treat the domain alignment objective and the classification objective as the meta-train and meta-test tasks in a meta-learning scheme. MetaAlign encourages both tasks to be optimized in a coordinated way, which maximizes the inner product of the gradients of the two tasks during training. Experimental results demonstrate the effectiveness of our proposed method on top of various alignment-based baseline approaches, for tasks of object classification and object detection. MetaAlign helps achieve the state-of-the-art performance.
****
Neural Prototype Trees for Interpretable Fine-Grained Image Recognition
Meike Nauta, Ron van Bree, Christin Seifert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14933-14943
Prototype-based methods use interpretable representations to address the black-box nature of deep learning models, in contrast to post-hoc explanation methods that only approximate such models. We propose the Neural Prototype Tree (ProtoTree), an intrinsically interpretable deep learning method for fine-grained image recognition. ProtoTree combines prototype learning with decision trees, and thus results in a globally interpretable model by design. Additionally, ProtoTree can

locally explain a single prediction by outlining a decision path through the tr
ee. Each node in our binary tree contains a trainable prototypical part. The pre
sence or absence of this learned prototype in an image determines the routing th
rough a node. Decision making is therefore similar to human reasoning: Does the
bird have a red throat? And an elongated beak? Then it's a hummingbird! We tune
the accuracy-interpretability trade-off using ensemble methods, pruning and bina
rizing. We apply pruning without sacrificing accuracy, resulting in a small tree
 with only 8 learned prototypes along a path to classify a bird from 200 species
. An ensemble of 5 ProtoTrees achieves competitive accuracy on the CUB-200- 2011
 and Stanford Cars data sets. Code is available at https://github.com/M-Nauta/Pr
otoTree.
****
Hardness Sampling for Self-Training Based Transductive Zero-Shot Learning
Liu Bo, Qiulei Dong, Zhanyi Hu; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 16499-16508
Transductive zero-shot learning (T-ZSL) which could alleviate the domain shift p
roblem in existing ZSL works, has received much attention recently. However, an
open problem in T-ZSL: how to effectively make use of unseen-class samples for t
raining, still remains. Addressing this problem, we first empirically analyze th
e roles of unseen-class samples with different degrees of hardness in the traini
ng process based on the uneven prediction phenomenon found in many ZSL methods,
resulting in three observations. Then, we propose two hardness sampling approach
es for selecting a subset of diverse and hard samples from a given unseen-class
dataset according to these observations. The first one identifies the samples ba
sed on the class-level frequency of the model predictions while the second enhan
ces the former by normalizing the class frequency via an approximate class prior
 estimated by an explored prior estimation algorithm. Finally, we design a new S
elf-Training framework with Hardness Sampling for T-ZSL, called STHS, where an a
rbitrary inductive ZSL method could be seamlessly embedded and it is iteratively
 trained with unseen-class samples selected by the hardness sampling approach. W
e introduce two typical ZSL methods into the STHS framework and extensive experi
ments demonstrate that the derived T-ZSL methods outperform many state-of-the-ar
t methods on three public benchmarks. Besides, we note that the unseen-class dat
aset is separately used for training in some existing transductive generalized Z
SL (T-GZSL) methods, which is not strict for a GZSL task. Hence, we suggest a mo
re strict T-GZSL data setting and establish a competitive baseline on this setti
ng by introducing the proposed STHS framework to T-GZSL.
****
Hilbert Sinkhorn Divergence for Optimal Transport
Qian Li, Zhichao Wang, Gang Li, Jun Pang, Guandong Xu; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3835-
3844
Sinkhorn divergence has become a very popular metric to compare probability dist
ributions in optimal transport. However, most works resort to Sinkhorn divergenc
e in Euclidean space, which greatly blocks their applications in complex data wi
th nonlinear structure. It is therefore of theoretical demand to empower Sinkhor
n divergence with the capability of capturing nonlinear structures. We propose a
 theoretical and computational framework to bridge this gap. In this paper, we e
xtend Sinkhorn divergence in Euclidean space to the reproducing kernel Hilbert s
pace, which we term "Hilbert Sinkhorn divergence" (HSD).In particular, we can us
e kernel matrices to derive a closed form expression of HSD that is proved to be
 a tractable convex optimization problem. We also prove several attractive stati
stical properties of the proposed HSD, i.e., strong consistency, asymptotic beha
vior and sample complexity. Empirically, our method yields state-of-the-art perf
ormances on image classification and topological data analysis.
****
The Multi-Temporal Urban Development SpaceNet Dataset
Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas We
ir, Ryan Lewis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2021, pp. 6398-6407

Satellite imagery analytics have numerous human development and disaster response applications, particularly when time series methods are involved. For example, quantifying population statistics is fundamental to 67 of the 231 United Nations Sustainable Development Goals Indicators, but the World Bank estimates that over 100 countries currently lack effective Civil Registration systems. To help address this deficit and develop novel computer vision methods for time series data, we present the Multi-Temporal Urban Development SpaceNet (MUDS, also known as SpaceNet 7) dataset. This open source dataset consists of medium resolution (4.0m) satellite imagery mosaics, which includes 24 images (one per month) covering >100 unique geographies, and comprises >40,000 km2 of imagery and exhaustive polygon labels of building footprints therein, totaling over 11M individual annotations. Each building is assigned a unique identifier (i.e. address), which permits tracking of individual objects over time. Label fidelity exceeds image resolution; this "omniscient labeling" is a unique feature of the dataset, and enables surprisingly precise algorithmic models to be crafted. We demonstrate methods to track building footprint construction (or demolition) over time, thereby directly assessing urbanization. Performance is measured with the newly developed SpaceNet Change and Object Tracking (SCOT) metric, which quantifies both object tracking as well as change detection. We demonstrate that despite the moderate resolution of the data, we are able to track individual building identifiers over time.
****
FBNetV3: Joint Architecture-Recipe Search Using Predictor Pretraining
Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, Joseph E. Gonzalez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16276-16285
Neural Architecture Search (NAS) yields state-of-the-art neural networks that outperform their best manually-designed counterparts. However, previous NAS methods search for architectures under one set of training hyper-parameters (i.e., a training recipe), overlooking superior architecture-recipe combinations. To address this, we present Neural Architecture-Recipe Search (NARS) to search both (a) architectures and (b) their corresponding training recipes, simultaneously. NARS utilizes an accuracy predictor that scores architecture and training recipes jointly, guiding both sample selection and ranking. Furthermore, to compensate for the enlarged search space, we leverage "free" architecture statistics (e.g., FLOP count) to pretrain the predictor, significantly improving its sample efficiency and prediction reliability. After training the predictor via constrained iterative optimization, we run fast evolutionary searches in just CPU minutes to generate architecture-recipe pairs for a variety of resource constraints, called FBNetV3. FBNetV3 makes up a family of state-of-the-art compact neural networks that outperform both automatically and manually-designed competitors. For example, FBNetV3 matches both EfficientNet and ResNeSt accuracy on ImageNet with up to 2.0x and 7.1x fewer FLOPs, respectively. Furthermore, FBNetV3 yields significant performance gains for downstream object detection tasks, improving mAP despite 18% fewer FLOPs and 34% fewer parameters than EfficientNet-based equivalents.
****
Intrinsic Image Harmonization
Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, Bing Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16367-16376
Compositing an image usually inevitably suffers from inharmony problem that is mainly caused by incompatibility of foreground and background from two different images with distinct surfaces and lights, corresponding to material-dependent and light-dependent characteristics, namely, reflectance and illumination intrinsic images, respectively. Therefore, we seek to solve image harmonization via separable harmonization of reflectance and illumination, i.e., intrinsic image harmonization. Our method is based on an autoencoder that disentangles composite image into reflectance and illumination for further separate harmonization. Specifically, we harmonize reflectance through material-consistency penalty, while harmo

nize illumination by learning and transferring light from background to foreground, moreover, we model patch relations between foreground and background of composite images in an inharmony-free learning way, to adaptively guide our intrinsic image harmonization. Both extensive experiments and ablation studies demonstrate the power of our method as well as the efficacy of each component. We also contribute a new challenging dataset for benchmarking illumination harmonization. Code and dataset are at https://github.com/zhenglab/IntrinsicHarmony.
****
L2M-GAN: Learning To Manipulate Latent Space Semantics for Facial Attribute Editing
Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, Tao Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2951-2960
A deep facial attribute editing model strives to meet two requirements: (1) attribute correctness -- the target attribute should correctly appear on the edited face image; (2) irrelevance preservation -- any irrelevant information (e.g., identity) should not be changed after editing. Meeting both requirements challenges the state-of-the-art works which resort to either spatial attention or latent space factorization. Specifically, the former assume that each attribute has well-defined local support regions; they are often more effective for editing a local attribute than a global one. The latter factorize the latent space of a fixed pretrained GAN into different attribute-relevant parts, but they cannot be trained end-to-end with the GAN, leading to sub-optimal solutions. To overcome these limitations, we propose a novel latent space factorization model, called L2M-GAN, which is learned end-to-end and effective for editing both local and global attributes. The key novel components are: (1) A latent space vector of the GAN is factorized into an attribute-relevant and irrelevant codes with an orthogonality constraint imposed to ensure disentanglement. (2) An attribute-relevant code transformer is learned to manipulate the attribute value; crucially, the transformed code are subject to the same orthogonality constraint. By forcing both the original attribute-relevant latent code and the edited code to be disentangled from any attribute-irrelevant code, our model strikes the perfect balance between attribute correctness and irrelevance preservation. Extensive experiments on CelebA-HQ show that our L2M-GAN achieves significant improvements over the state-of-the-arts.
****
IIRC: Incremental Implicitly-Refined Classification
Mohamed Abdelsalam, Mojtaba Faramarzi, Shagun Sodhani, Sarath Chandar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11038-11047
We introduce the 'Incremental Implicitly-Refined Classification (IIRC)' setup, an extension to the class incremental learning setup where the incoming batches of classes have two granularity levels. i.e., each sample could have a high-level (coarse) label like 'bear' and a low-level (fine) label like 'polar bear'. Only one label is provided at a time, and the model has to figure out the other label if it has already learned it. This setup is more aligned with real-life scenarios, where a learner usually interacts with the same family of entities multiple times, discovers more granularity about them, while still trying not to forget previous knowledge. Moreover, this setup enables evaluating models for some important lifelong learning challenges that cannot be easily addressed under the existing setups. These challenges can be motivated by the example "if a model was trained on the class 'bear' in one task and on 'polar bear' in another task, will it forget the concept of 'bear', will it rightfully infer that a 'polar bear' is still a 'bear'? and will it wrongfully associate the label of 'polar bear' to other breeds of 'bear'?". We develop a standardized benchmark that enables evaluating models on the IIRC setup. We evaluate several state-of-the-art lifelong learning algorithms and highlight their strengths and limitations. For example, distillation-based methods perform relatively well but are prone to incorrectly predicting too many labels per image. We hope that the proposed setup, along with the benchmark, would provide a meaningful problem setting to the practitioners.

****
Learning To Fuse Asymmetric Feature Maps in Siamese Trackers
Wencheng Han, Xingping Dong, Fahad Shahbaz Khan, Ling Shao, Jianbing Shen; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2021, pp. 16570-16580
Recently, Siamese-based trackers have achieved promising performance in visual t
racking. Most recent Siamese-based trackers typically employ a depth-wise cross-
correlation (DW-XCorr) to obtain multi-channel correlation information from the
two feature maps (target and search region). However, DW-XCorr has several limit
ations within Siamese-based tracking: it can easily be fooled by distractors, ha
s fewer activated channels, and provides weak discrimination of object boundarie
s. Further, DW-XCorr is a handcrafted parameter-free module and cannot fully ben
efit from offline learning on large-scale data. We propose a learnable module, c
alled the asymmetric convolution (ACM), which learns to better capture the seman
tic correlation information in offline training on large-scale data. Different f
rom DW-XCorr and its predecessor(XCorr), which regard a single feature map as th
e convolution kernel, our ACM decomposes the convolution operation on a concaten
ated feature map into two mathematically equivalent operations, thereby avoiding
 the need for the feature maps to be of the same size (width and height)during c
oncatenation. Our ACM can incorporate useful prior information, such as bounding
-box size, with standard visual features. Furthermore, ACM can easily be integra
ted into existing Siamese trackers based on DW-XCorror XCorr. To demonstrate its
 generalization ability, we integrate ACM into three representative trackers: Si
amFC, SiamRPN++, and SiamBAN. Our experiments reveal the benefits of the propose
d ACM, which outperforms existing methods on six tracking benchmarks. On the LaS
OT test set, our ACM-based tracker obtains a significant improvement of 5.8% in
terms of success (AUC), over the baseline.
****
Generalizing to the Open World: Deep Visual Odometry With Online Adaptation
Shunkai Li, Xin Wu, Yingdian Cao, Hongbin Zha; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13184-13193
Despite learning-based visual odometry (VO) has shown impressive results in rece
nt years, the pretrained networks may easily collapse in unseen environments. Th
e large domain gap between training and testing data makes them difficult to gen
eralize to new scenes. In this paper, we propose an online adaptation framework
for deep VO with the assistance of scene-agnostic geometric computations and Bay
esian inference. In contrast to learning-based pose estimation, our method solve
s pose from optical flow and depth while the single-view depth estimation is con
tinuously improved with new observations by online learned uncertainties. Meanwh
ile, an online learned photometric uncertainty is used for further depth and pos
e optimization by a differentiable Gauss-Newton layer. Our method enables fast a
daptation of deep VO networks to unseen environments in a self-supervised manner
. Extensive experiments including Cityscapes to KITTI and outdoor KITTI to indoo
r TUM demonstrate that our method achieves state-of-the-art generalization abili
ty among self-supervised VO methods.
****
PQA: Perceptual Question Answering
Yonggang Qi, Kai Zhang, Aneeshan Sain, Yi-Zhe Song; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12056-12
064
Perceptual organization remains one of the very few established theories on the
human visual system. It underpinned many pre-deep seminal works on segmentation
and detection, yet research has seen a rapid decline since the preferential shif
t to learning deep models. Of the limited attempts, most aimed at interpreting c
omplex visual scenes using perceptual organizational rules. This has however bee
n proven to be sub-optimal, since models were unable to effectively capture the
visual complexity in real-world imagery. In this paper, we rejuvenate the study
of perceptual organization, by advocating two positional changes: (i) we examine
 purposefully generated synthetic data, instead of complex real imagery, and (ii
) we ask machines to synthesize novel perceptually-valid patterns, instead of ex

plaining existing data. Our overall answer lies with the introduction of a novel visual challenge -- the challenge of perceptual question answering (PQA). Upon observing example perceptual question-answer pairs, the goal for PQA is to solve similar questions by generating answers entirely from scratch (see Figure 1). Our first contribution is therefore the first dataset of perceptual question-answer pairs, each generated specifically for a particular Gestalt principle. We then borrow insights from human psychology to design an agent that casts perceptual organization as a self-attention problem, where a proposed grid-to-grid mapping network directly generates answer patterns from scratch. Experiments show our agent to outperform a selection of naive and strong baselines. A human study however indicates that ours uses astronomically more data to learn when compared to an average human, necessitating future research (with or without our dataset).
****
Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink
Ranjie Duan, Xiaofeng Mao, A. K. Qin, Yuefeng Chen, Shaokai Ye, Yuan He, Yun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16062-16071
Though it is well known that the performance of deep neural networks (DNNs) degrades under certain light conditions, there exists no study on the threats of light beams emitted from some physical source as adversarial attacker on DNNs in a real-world scenario. In this work, we show by simply using a laser beam that DNNs are easily fooled. To this end, we propose a novel attack method called Adversarial Laser Beam (AdvLB), which enables manipulation of laser beam's physical parameters to perform adversarial attack. Experiments demonstrate the effectiveness of our proposed approach in both digital- and physical-settings. We further empirically analyze the evaluation results and reveal that the proposed laser beam attack may lead to some interesting prediction errors of the state-of-the-art DNNs. We envisage that the proposed AdvLB method enriches the current family of adversarial attacks and builds the foundation for future robustness studies for light.
****
Robust Point Cloud Registration Framework Based on Deep Graph Matching
Kexue Fu, Shaolei Liu, Xiaoyuan Luo, Manning Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8893-8902
3D point cloud registration is a fundamental problem in computer vision and robotics. Recently, learning-based point cloud registration methods have made great progress. However, these methods are sensitive to outliers, which lead to more incorrect correspondences. In this paper, we propose a novel deep graph matching-based framework for point cloud registration. Specifically, we first transform point clouds into graphs and extract deep features for each point. Then, we develop a module based on deep graph matching to calculate a soft correspondence matrix. By using graph matching, not only the local geometry of each point but also its structure and topology in a larger range are considered in establishing correspondences, so that more correct correspondences are found. We train the network with a loss directly defined on the correspondences, and in the test stage the soft correspondences are transformed into hard one-to-one correspondences so that registration can be performed by singular value decomposition. Furthermore, we introduce a transformer-based method to generate edges for graph construction, which further improves the quality of the correspondences. Extensive experiments on registering clean, noisy, partial-to-partial and unseen category point clouds show that the proposed method achieves state-of-the-art performance. The code will be made publicly available at https://github.com/fukexue/RGM.
****
Dense Contrastive Learning for Self-Supervised Visual Pre-Training
Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, Lei Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3024-3033
To date, most existing self-supervised learning methods are designed and optimized for image classification. These pre-trained models can be sub-optimal for dense prediction tasks due to the discrepancy between image-level prediction and pi

xel-level prediction. To fill this gap, we aim to design an effective, dense sel
f-supervised learning method that directly works at the level of pixels (or loca
l features) by taking into account the correspondence between local features. We
 present dense contrastive learning, which implements self-supervised learning b
y optimizing a pairwise contrastive (dis)similarity loss at the pixel level betw
een two views of input images. Compared to the baseline method MoCo-v2, our meth
od introduces negligible computation overhead (only <1% slower), but demonstrate
s consistently superior performance when transferring to downstream dense predic
tion tasks including object detection, semantic segmentation and instance segmen
tation; and outperforms the state-of-the-art methods by a large margin. Specific
ally, over the strong MoCo-v2 baseline, our method achieves significant improvem
ents of 2.0% AP on PASCAL VOC object detection, 1.1% AP on COCO object detection
, 0.9% AP on COCO instance segmentation, 3.0% mIoU on PASCAL VOC semantic segmen
tation and 1.8% mIoU on Cityscapes semantic segmentation. Code and models are av
ailable at: https://git.io/DenseCL
****
Birds of a Feather: Capturing Avian Shape Models From Images
Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, Marc Badger; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
 14739-14749
Animals are diverse in shape, but building a deformable shape model for a new sp
ecies is not always possible due to the lack of 3D data. We present a method to
capture new species using an articulated template and images of that species. In
 this work, we focus mainly on birds. Although birds represent almost twice the
number of species as mammals, no accurate shape model is available. To capture a
 novel species, we first fit the articulated template to each training sample. B
y disentangling pose and shape, we learn a shape space that captures variation b
oth among species and within each species from image evidence. We learn models o
f multiple species from the CUB dataset, and contribute new species-specific and
 multi-species shape models that are useful for downstream reconstruction tasks.
 Using a low-dimensional embedding, we show that our learned 3D shape space bett
er reflects the phylogenetic relationships among birds than learned perceptual f
eatures.
****
Learning Temporal Consistency for Low Light Video Enhancement From Single Images
Fan Zhang, Yu Li, Shaodi You, Ying Fu; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4967-4976
Single image low light enhancement is an important task and it has many practica
l applications. Most existing methods adopt a single image approach. Although th
eir performance is satisfying on a static single image, we found, however, they
suffer serious temporal instability when handling low light videos. We notice th
e problem is because existing data-driven methods are trained from single image
pairs where no temporal information is available. Unfortunately, training from r
eal temporally consistent data is also problematic because it is impossible to c
ollect pixel-wisely paired low and normal light videos under controlled environm
ents in large scale and diversities with noise of identical statistics. In this
paper, we propose a novel method to enforce the temporal stability in low light
video enhancement with only static images. The key idea is to learn and infer mo
tion field (optical flow) from a single image and synthesize short range video s
equences. Our strategy is general and can extend to large scale datasets directl
y. Based on this idea, we propose our method which can infer motion prior for si
ngle image low light video enhancement and enforce temporal consistency. Rigorou
s experiments and user study demonstrate the state-of-the-art performance of our
 proposed method. Our code and model will be publicly available at https://githu
b.com/zkawfanx/StableLLVE.
****
Brain Image Synthesis With Unsupervised Multivariate Canonical CSCl4Net
Yawen Huang, Feng Zheng, Danyang Wang, Weilin Huang, Matthew R. Scott, Ling Shao
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2021, pp. 5881-5890

Recent advances in neuroscience have highlighted the effectiveness of multi-moda
l medical data for investigating certain pathologies and understanding human cog
nition. However, obtaining full sets of different modalities is limited by vario
us factors, such as long acquisition times, high examination costs and artifact
suppression. In addition, the complexity, high dimensionality and heterogeneity
of neuroimaging data remains another key challenge in leveraging existing random
ized scans effectively, as data of the same modality is often measured different
ly by different machines. There is a clear need to go beyond the traditional ima
ging-dependent process and synthesize anatomically specific target-modality data
 from a source input. In this paper, we propose to learn dedicated features that
 cross both intre- and intra-modal variations using a novel CSCl_4Net. Through a
n initial unification of intra-modal data in the feature maps and multivariate c
anonical adaptation, CSCl_4Net facilitates feature-level mutual transformation.
The positive definite Riemannian manifold-penalized data fidelity term further e
nables CSCl_4Net to reconstruct missing measurements according to transformed fe
atures. Finally, the maximization l_4-norm boils down to a computationally effic
ient optimization problem. Extensive experiments validate the ability and robust
ness of our CSCl_4Net compared to the state-of-the-art methods on multiple datas
ets.
****
Inverse Simulation: Reconstructing Dynamic Geometry of Clothed Humans via Optima
l Control
Jingfan Guo, Jie Li, Rahul Narain, Hyun Soo Park; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14698-1470
7
This paper studies the problem of inverse cloth simulation---to estimate shape a
nd time-varying poses of the underlying body that generates physically plausible
 cloth motion, which matches to the point cloud measurements on the clothed huma
ns. A key innovation is to represent the dynamics of the cloth geometry using a
dynamical system that is controlled by the body states (shape and pose). This al
lows us to express the cloth motion as a resultant of external (skin friction an
d gravity) and internal (elasticity) forces. Inspired by the theory of optimal c
ontrol, we optimize the body states such that the simulated cloth motion is matc
hed to the point cloud measurements, and the analytic gradient of the simulator
is back-propagated to update the body states. We propose a cloth relaxation sche
me to initialize the cloth state, which ensures the physical validity. Our metho
d produces physically plausible and temporally smooth cloth and body movements t
hat are faithful to the measurements, and shows superior performance compared to
 the existing methods. As a byproduct, the stress and strain that are applied to
 the body and clothes can be recovered.
****
Rotation Equivariant Siamese Networks for Tracking
Deepak K. Gupta, Devanshu Arya, Efstratios Gavves; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12362-123
71
Rotation is among the long prevailing, yet still unresolved, hard challenges enc
ountered in visual object tracking. The existing deep learning-based tracking al
gorithms use regular CNNs that are inherently translation equivariant, but not d
esigned to tackle rotations. In this paper, we first demonstrate that in the pre
sence of rotation instances in videos, the performance of existing trackers is s
everely affected. To circumvent the adverse effect of rotations, we present rota
tion-equivariant Siamese networks (RE-SiamNets), built through the use of group-
equivariant convolutional layers comprising steerable filters. SiamNets allow es
timating the change in orientation of the object in an unsupervised manner, ther
eby facilitating its use in relative 2D pose estimation as well. We further show
 that this change in orientation can be used to impose an additional motion cons
traint in Siamese tracking through imposing restriction on the change in orienta
tion between two consecutive frames. For benchmarking, we present Rotation Track
ing Benchmark (RTB), a dataset comprising a set of videos with rotation instance
s. Through experiments on two popular Siamese architectures, we show that RE-Sia

mNets handle the problem of rotation very well and outperform their regular counterparts. Further, RE-SiamNets can accurately estimate the relative change in pose of the target in an unsupervised fashion, namely the in-plane rotation the target has sustained with respect to the reference frame.
****
Learning Decision Trees Recurrently Through Communication
Stephan Alaniz, Diego Marcos, Bernt Schiele, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13518-13527
Integrated interpretability without sacrificing the prediction accuracy of decision making algorithms has the potential of greatly improving their value to the user. Instead of assigning a label to an image directly, we propose to learn iterative binary sub-decisions, inducing sparsity and transparency in the decision making process. The key aspect of our model is its ability to build a decision tree whose structure is encoded into the memory representation of a Recurrent Neural Network jointly learned by two models communicating through message passing. In addition, our model assigns a semantic meaning to each decision in the form of binary attributes, providing concise, semantic and relevant rationalizations to the user. On three benchmark image classification datasets, including the large-scale ImageNet, our model generates human interpretable binary decision sequences explaining the predictions of the network while maintaining state-of-the-art accuracy.
****
PatchmatchNet: Learned Multi-View Patchmatch Stereo
Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14194-14203
We present PatchmatchNet, a novel and learnable cascade formulation of Patchmatch for high-resolution multi-view stereo. With high computation speed and low memory requirement, PatchmatchNet can process higher resolution imagery and is more suited to run on resource limited devices than competitors that employ 3D cost volume regularization. For the first time we introduce an iterative multi-scale Patchmatch in an end-to-end trainable architecture and improve the Patchmatch core algorithm with a novel and learned adaptive propagation and evaluation scheme for each iteration. Extensive experiments show a very competitive performance and generalization for our method on DTU, Tanks & Temples and ETH3D, but at a significantly higher efficiency than all existing top-performing models: at least two and a half times faster than state-of-the-art methods with twice less memory usage. Code is available at https://github.com/FangjinhuaWang/PatchmatchNet.
****
Instance Level Affinity-Based Transfer for Unsupervised Domain Adaptation
Astuti Sharma, Tarun Kalluri, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5361-5371
Domain adaptation deals with training models using large scale labeled data from a specific source domain and then adapting the knowledge to certain target domains that have few or no labels. Many prior works learn domain agnostic feature representations for this purpose using a global distribution alignment objective which does not take into account the finer class specific structure in the source and target domains. We address this issue in our work and propose an instance affinity based criterion for source to target transfer during adaptation, called ILA-DA. We first propose a reliable and efficient method to extract similar and dissimilar samples across source and target, and utilize a multi-sample contrastive loss to drive the domain alignment process. ILA-DA simultaneously accounts for intra-class clustering as well as inter-class separation among the categories, resulting in less noisy classifier boundaries, improved transferability and increased accuracy. We verify the effectiveness of ILA-DA by observing consistent improvements in accuracy over popular domain adaptation approaches on a variety of benchmark datasets and provide insights into the proposed alignment approach. Code will be made publicly available at https://github.com/astuti/ILA-DA.
****

COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction
Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11174-11183
In this paper, we study two challenging problems in incomplete multi-view clustering analysis, namely, i) how to learn an informative and consistent representation among different views without the help of labels and ii) how to recover the missing views from data. To this end, we propose a novel objective that incorporates representation learning and data recovery into a unified framework from the view of information theory. To be specific, the informative and consistent representation is learned by maximizing the mutual information across different views through contrastive learning, and the missing views are recovered by minimizing the conditional entropy of different views through dual prediction. To the best of our knowledge, this could be the first work to provide a theoretical framework that unifies the consistent representation learning and cross-view data recovery. Extensive experimental results show the proposed method remarkably outperforms 10 competitive multi-view clustering methods on four challenging datasets. The code is available at https://pengxi.me.
****
Image-to-Image Translation via Hierarchical Style Disentanglement
Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8639-8648
Recently, image-to-image translation has made significant progress in achieving both multi-label (i.e., translation conditioned on different labels) and multi-style (i.e., generation with diverse styles) tasks. However, due to the unexplored independence and exclusiveness in the labels, existing endeavors are defeated by involving uncontrolled manipulations to the translation results. In this paper, we propose Hierarchical Style Disentanglement (HiSD) to address this issue. Specifically, we organize the labels into a hierarchical tree structure, in which independent tags, exclusive attributes, and disentangled styles are allocated from top to bottom. Correspondingly, a new translation process is designed to adapt the above structure, in which the styles are identified for controllable translations. Both qualitative and quantitative results on the CelebA-HQ dataset verify the ability of the proposed HiSD. We hope our method will serve as a solid baseline and provide fresh insights with the hierarchically organized annotations for future research in image-to-image translation. The code will be released.
****
What Can Style Transfer and Paintings Do for Model Robustness?
Hubert Lin, Mitchell van Zuijlen, Sylvia C. Pont, Maarten W.A. Wijntjes, Kavita Bala; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11028-11037
A common strategy for improving model robustness is through data augmentations. Data augmentations encourage models to learn desired invariances, such as invariance to horizontal flipping or small changes in color. Recent work has shown that arbitrary style transfer can be used as a form of data augmentation to encourage invariance to textures by creating painting-like images from photographs. However, a stylized photograph is not quite the same as an artist-created painting. Artists depict perceptually meaningful cues in paintings so that humans can recognize salient components in scenes, an emphasis which is not enforced in style transfer. Therefore, we study how style transfer and paintings differ in their impact on model robustness. First, we investigate the role of paintings as style images for stylization-based data augmentation. We find that style transfer functions well even without paintings as style images. Second, we show that learning from paintings as a form of perceptual data augmentation can improve model robustness. Finally, we investigate the invariances learned from stylization and from paintings, and show that models learn different invariances from these differing forms of data. Our results provide insights into how stylization improves model robustness, and provide evidence that artist-created paintings can be a valuable source of data for model robustness.

****
Taming Transformers for High-Resolution Image Synthesis
Patrick Esser, Robin Rombach, Bjorn Ommer; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12873-12883
Designed to learn long-range interactions on sequential data, transformers conti
nue to show state-of-the-art results on a wide variety of tasks. In contrast to
CNNs, they contain no inductive bias that prioritizes local interactions. This m
akes them expressive, but also computationally infeasible for long sequences, su
ch as high-resolution images. We demonstrate how combining the effectiveness of
the inductive bias of CNNs with the expressivity of transformers enables them to
 model and thereby synthesize high-resolution images. We show how to (i) use CNN
s to learn a context-rich vocabulary of image constituents, and in turn (ii) uti
lize transformers to efficiently model their composition within high-resolution
images. Our approach is readily applied to conditional synthesis tasks, where bo
th non-spatial information, such as object classes, and spatial information, suc
h as segmentations, can control the generated image. In particular, we present t
he first results on semantically-guided synthesis of megapixel images with trans
formers.
****
Learning the Predictability of the Future
Didac Suris, Ruoshi Liu, Carl Vondrick; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12607-12617
We introduce a framework for learning from unlabeled video what is predictable i
n the future. Instead of committing up front to features to predict, our approac
h learns from data which features are predictable. Based on the observation that
 hyperbolic geometry naturally and compactly encodes hierarchical structure, we
propose a predictive model in hyperbolic space. When the model is most confident
, it will predict at a concrete level of the hierarchy, but when the model is no
t confident, it learns to automatically select a higher level of abstraction. Ex
periments on two established datasets show the key role of hierarchical represen
tations for action prediction. Although our representation is trained with unlab
eled video, visualizations show that action hierarchies emerge in the representa
tion.
****
Multiple Instance Captioning: Learning Representations From Histopathology Textb
ooks and Articles
Jevgenij Gamper, Nasir Rajpoot; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 16549-16559
We present ARCH, a computational pathology (CP) multiple instance captioning dat
aset to facilitate dense supervision of CP tasks. Existing CP datasets focus on
narrow tasks; ARCH on the other hand contains dense diagnostic and morphological
 descriptions for a range of stains, tissue types and pathologies. Using intrins
ic dimensionality estimation, we show that ARCH is the only CP dataset to (ARCH-
)rival its computer vision analog MS-COCO Captions. We conjecture that an encode
r pre-trained on dense image captions learns transferable representations for mo
st CP tasks. We support the conjecture with evidence that ARCH representation tr
ansfers to a variety of pathology sub-tasks better than ImageNet features or rep
resentations obtained via self-supervised or multi-task learning on pathology im
ages alone. We release our best model and invite other researchers to test it on
 their CP tasks.
****
Beyond Max-Margin: Class Margin Equilibrium for Few-Shot Object Detection
Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, Qixiang Ye; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
1, pp. 7363-7372
Few-shot object detection has made encouraging progress by reconstructing novel
class objects using the feature representation learned upon a set of base classe
s. However, an implicit contradiction about reconstruction and classification is
 unfortunately ignored. On the one hand, to precisely reconstruct novel classes,
 the distributions of base classes should be close to those of novel classes (mi

n-margin). On the other hand, to perform accurate classification, the distributi ons of either two classes must be far away from each other (max-margin). In this paper, we propose a class margin equilibrium (CME) approach, with the aim to op timize both feature space partition and novel class reconstruction in a systemat ic way. CME first converts the few-shot detection problem to the few-shot classi fication problem by using a fully connection layer to decouple localization feat ures. CME then reserves adequate margin space for novel classes by introducing s imple-yet-effective class margin loss during feature learning. Finally, CME purs ues margin equilibrium by disturbing the features of novel class instances in an adversarial min-max fashion. Experiments on Pascal VOC and MS-COCO datasets sho w that CME improves two baseline detectors (up to 5% in average), achieving new state-of-the-art performance.
****

Consistent Instance False Positive Improves Fairness in Face Recognition
Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, Zhen Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2021, pp. 578-586
Demographic bias is a significant challenge in practical face recognition system s. Several methods have been proposed to reduce the bias, which rely on accurate demographic annotations. However, such annotations are usually not available in real scenarios. Moreover, these methods are explicitly designed for a specific demographic group divided by a predefined attribute, which is typically not gene ral across different demographic groups divided by various attributes, such as r ace, gender, and age. In this paper, we propose a false positive rate penalty lo ss, which mitigates face recognition bias by increasing the consistency of insta nce false positive rate (FPR). Specifically, we first define the instance FPR as the ratio between the number of the non-target similarities above a unified thr eshold and the total number of the non-target similarities. The unified threshol d is estimated for a given total FPR. Then, we introduce an additional false pos itive penalty term into the softmaxbased losses to promote the consistency of in stance FPRs. Compared with the previous debiasing methods, our method requires n o demographic annotations and can mitigate the bias across demographic groups di vided by various kinds of attribute which are no need to be predefined in traini ng. Extensive experimental results on popular benchmarks demonstrate the superio rity of our method over state-of-the art competitors.
****

Learning Dynamic Network Using a Reuse Gate Function in Semi-Supervised Video Ob ject Segmentation
Hyojin Park, Jayeon Yoo, Seohyeong Jeong, Ganesh Venkatesh, Nojun Kwak; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2021, pp. 8405-8414
Current state-of-the-art approaches for Semi-supervised Video Object Segmentatio n (Semi-VOS) propagates information from previous frames to generate segmentatio n mask for the current frame. This results in high-quality segmentation across c hallenging scenarios such as changes in appearance and occlusion. But it also le ads to unnecessary computations for stationary or slow-moving objects where the change across frames is minimal. In this work, we exploit this observation by us ing temporal information to quickly identify frames with minimal change and skip the heavyweight mask generation step. To realize this efficiency, we propose a novel dynamic network that estimates change across frames and decides which path -- computing a full network or reusing previous frame's feature -- to choose de pending on the expected similarity. Experimental results show that our approach significantly improves inference speed without much accuracy degradation on chal lenging Semi-VOS datasets -- DAVIS 16, DAVIS 17, and YouTube-VOS. Furthermore, o ur approach can be applied to multiple Semi-VOS methods demonstrating its genera lity. The code is available in https://github.com/HYOJINPARK/Reuse VOS.
****

RaScaNet: Learning Tiny Models by Raster-Scanning Images
Jaehyoung Yoo, Dongwook Lee, Changyong Son, Sangil Jung, ByungIn Yoo, Changkyu C hoi, Jae-Joon Han, Bohyung Han; Proceedings of the IEEE/CVF Conference on Comput

er Vision and Pattern Recognition (CVPR), 2021, pp. 13673-13682
Deploying deep convolutional neural networks on ultra-low power systems is chall enging due to the extremely limited resources. Especially, the memory becomes a bottleneck as the systems put a hard limit on the size of on-chip memory. Becaus e peak memory explosion in the lower layers is critical even in tiny models, the size of an input image should be reduced with sacrifice in accuracy. To overcom e this drawback, we propose a novel Raster-Scanning Network, named RaScaNet, ins pired by raster-scanning in image sensors. RaScaNet reads only a few rows of pix els at a time using a convolutional neural network and then sequentially learns the representation of the whole image using a recurrent neural network. The prop osed method operates on an ultra-low power system without input size reduction; it requires 15.9-24.3x smaller peak memory and 5.3-12.9x smaller weight memory t han the state-of-the-art tiny models. Moreover, RaScaNet fully exploits on-chip SRAM and cache memory of the system as the sum of the peak memory and the weight memory does not exceed 60 KB, improving the power efficiency of the system. In our experiments, we demonstrate the binary classification performance of RaScaNe t on Visual Wake Words and Pascal VOC datasets.
****
AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning
Madeleine Grunde-McLaughlin, Ranjay Krishna, Maneesh Agrawala; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 11287-11297
Visual events are a composition of temporal actions involving actors spatially i nteracting with objects. When developing computer vision models that can reason about compositional spatio-temporal events, we need benchmarks that can analyze progress and uncover shortcomings. Existing video question answering benchmarks are useful, but they often conflate multiple sources of error into one accuracy metric and have strong biases that models can exploit, making it difficult to pi npoint model weaknesses. We present Action Genome Question Answering (AGQA), a n ew benchmark for compositional spatio-temporal reasoning. AGQA contains 192M unb alanced question answer pairs for 9.6K videos. We also provide a balanced subset of 3.9M question answer pairs, 3 orders of magnitude larger than existing bench marks, that minimizes bias by balancing the answer distributions and types of qu estion structures. Although human evaluators marked 86.02% of our question-answe r pairs as correct, the best model achieves only 47.74% accuracy. In addition, A GQA introduces multiple training/test splits to test for various reasoning abili ties, including generalization to novel compositions, to indirect references, an d to more compositional steps. Using AGQA, we evaluate modern visual reasoning s ystems, demonstrating that the best models barely perform better than non-visual baselines exploiting linguistic biases and that none of the existing models gen eralize to novel compositions unseen during training.
****
Exploring intermediate representation for monocular vehicle pose estimation
Shichao Li, Zengqiang Yan, Hongyang Li, Kwang-Ting Cheng; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 18 73-1883
We present a new learning-based framework to recover vehicle pose in SO(3) from a single RGB image. In contrast to previous works that map local appearance to o bservation angles, we explore a progressive approach by extracting meaningful In termediate Geometrical Representations (IGRs) to estimate egocentric vehicle ori entation. This approach features a deep model that transforms perceived intensit ies to IGRs, which are mapped to a 3D representation encoding object orientation in the camera coordinate system. Core problems are what IGRs to use and how to learn them more effectively. We answer the former question by designing IGRs bas ed on an interpolated cuboid that derives from primitive 3D annotation readily. The latter question motivates us to incorporate geometry knowledge with a new lo ss function based on a projective invariant. This loss function allows unlabeled data to be used in the training stage to improve representation learning. Witho ut additional labels, our system outperforms previous monocular RGB-based method s for joint vehicle detection and pose estimation on the KITTI benchmark, achiev

ing performance even comparable to stereo methods. Code and pre-trained models a re available at this HTTPS URL.
****

Shallow Feature Matters for Weakly Supervised Object Localization
Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S. Kevin Zhou, Shuguang Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5993-6001
Weakly supervised object localization (WSOL) aims to localize objects by only utilizing image-level labels. Class activation maps (CAMs) are the commonly used features to achieve WSOL. However, previous CAM-based methods did not take full advantage of the shallow features, despite their importance for WSOL. Because shallow features are easily buried in background noise through conventional fusion. In this paper, we propose a simple but effective Shallow feature-aware Pseudo supervised Object Localization (SPOL) model for accurate WSOL, which makes the utmost of low-level features embedded in shallow layers. In practice, our SPOL model first generates the CAMs through a novel element-wise multiplication of shallow and deep feature maps, which filters the background noise and generates sharper boundaries robustly. Besides, we further propose a general class-agnostic segmentation model to achieve the accurate object mask, by only using the initial CAMs as the pseudo label without any extra annotation. Eventually, a bounding box extractor is applied to the object mask to locate the target. Experiments verify that our SPOL outperforms the state-of-the-art on both CUB-200 and ImageNet-1K benchmarks, achieving 93.44% and 67.15% (i.e., 3.93% and 2.13% improvement) Top-5 localization accuracy, respectively.
****

Capturing Omni-Range Context for Omnidirectional Segmentation
Kailun Yang, Jiaming Zhang, Simon Reiss, Xinxin Hu, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1376-1386
Convolutional Networks (ConvNets) excel at semantic segmentation and have become a vital component for perception in autonomous driving. Enabling an all-encompassing view of street-scenes, omnidirectional cameras present themselves as a perfect fit in such systems. Most segmentation models for parsing urban environments operate on common, narrow Field of View (FoV) images. Transferring these models from the domain they were designed for to 360-degree perception, their performance drops dramatically, e.g., by an absolute 30.0% (mIoU) on established test-beds. To bridge the gap in terms of FoV and structural distribution between the imaging domains, we introduce Efficient Concurrent Attention Networks (ECANets), directly capturing the inherent long-range dependencies in omnidirectional imagery. In addition to the learned attention-based contextual priors that can stretch across 360-degree images, we upgrade model training by leveraging multi-source and omni-supervised learning, taking advantage of both: Densely labeled and unlabeled data originating from multiple datasets. To foster progress in panoramic image segmentation, we put forward and extensively evaluate models on Wild PAnoramic Semantic Segmentation (WildPASS), a dataset designed to capture diverse scenes from all around the globe. Our novel model, training regimen and multi-source prediction fusion elevate the performance (mIoU) to new state-of-the-art results on the public PASS (60.2%) and the fresh WildPASS (69.0%) benchmarks.
****

PLADE-Net: Towards Pixel-Level Accuracy for Self-Supervised Single-View Depth Estimation With Neural Positional Encoding and Distilled Matting Loss
Juan Luis Gonzalez, Munchurl Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6851-6860
In this paper, we propose a self-supervised single-view pixel-level accurate depth estimation network, called PLADE-Net. The PLADE-Net is the first work that shows unprecedented accuracy levels, exceeding 95% in terms of the \delta^1 metric on the challenging KITTI dataset. Our PLADE-Net is based on a new network architecture with neural positional encoding and a novel loss function that borrows from the closed-form solution of the matting Laplacian to learn pixel-level accurate depth estimation from stereo images. Neural positional encoding allows our P

LADE-Net to obtain more consistent depth estimates by letting the network reason about location-specific image properties such as lens and projection distortions. Our novel distilled matting Laplacian loss allows our network to predict sharp depths at object boundaries and more consistent depths in highly homogeneous regions. Our proposed method outperforms all previous self-supervised single-view depth estimation methods by a large margin on the challenging KITTI dataset, with unprecedented levels of accuracy. Furthermore, our PLADE-Net, naively extended for stereo inputs, outperforms the most recent self-supervised stereo methods, even without any advanced blocks like 1D correlations, 3D convolutions, or spatial pyramid pooling. We present extensive ablation studies and experiments that support our method's effectiveness on the KITTI, CityScapes, and Make3D datasets.
****
Reciprocal Landmark Detection and Tracking With Extremely Few Annotations
Jianzhe Lin, Ghazal Sahebzamani, Christina Luong, Fatemeh Taheri Dezaki, Mohammad Jafari, Purang Abolmaesumi, Teresa Tsang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15170-15179
Localization of anatomical landmarks to perform two-dimensional measurements in echocardiography is part of routine clinical workflow in cardiac disease diagnosis. Automatic localization of those landmarks is highly desirable to improve workflow and reduce interobserver variability. Training a machine learning framework to perform such localization is hindered given the sparse nature of gold standard labels; only few percent of cardiac cine series frames are normally manually labeled for clinical use. In this paper, we propose a new end-to-end reciprocal detection and tracking model that is specifically designed to handle the sparse nature of echocardiography labels. The model is trained using few annotated frames across the entire cardiac cine sequence to generate consistent detection and tracking of landmarks, and an adversarial training for the model is proposed to take advantage of these annotated frames. The superiority of the proposed reciprocal model is demonstrated using a series of experiments.
****
Practical Single-Image Super-Resolution Using Look-Up Table
Younghyun Jo, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 691-700
A number of super-resolution (SR) algorithms from interpolation to deep neural networks (DNN) have emerged to restore or create missing details of the input low-resolution image. As mobile devices and display hardware develops, the demand for practical SR technology has increased. Current state-of-the-art SR methods are based on DNNs for better quality. However, they are feasible when executed by using a parallel computing module (e.g. GPUs), and have been difficult to apply to general uses such as end-user software, smartphones, and televisions. To this end, we propose an efficient and practical approach for the SR by adopting look-up table (LUT). We train a deep SR network with a small receptive field and transfer the output values of the learned deep model to the LUT. At test time, we retrieve the precomputed HR output values from the LUT for query LR input pixels. The proposed method can be performed very quickly because it does not require a large number of floating point operations. Experimental results show the efficiency and the effectiveness of our method. Especially, our method runs faster while showing better quality compared to bicubic interpolation.
****
Removing the Background by Adding the Background: Towards Background Robust Self-Supervised Video Representation Learning
Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J. Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, Xing Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11804-11813
Self-supervised learning has shown great potentials in improving the video representation ability of deep neural networks by getting supervision from the data itself. However, some of the current methods tend to cheat from the background, i.e., the prediction is highly dependent on the video background instead of the motion, making the model vulnerable to background changes. To mitigate the model

reliance towards the background, we propose to remove the background impact by adding the background. That is, given a video, we randomly select a static frame and add it to every other frames to construct a distracting video sample. Then we force the model to pull the feature of the distracting video and the feature of the original video closer, so that the model is explicitly restricted to resist the background influence, focusing more on the motion changes. We term our method as Background Erasing (BE). It is worth noting that the implementation of our method is so simple and neat and can be added to most of the SOTA methods without much efforts. Specifically, BE brings 16.4% and 19.1% improvements with MoCo on the severely biased datasets UCF101 and HMDB51, and 14.5% improvement on the less biased dataset Diving48.
****

GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation

Gu Wang, Fabian Manhardt, Federico Tombari, Xiangyang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16611-16621

6D pose estimation from a single RGB image is a fundamental task in computer vision. The current top-performing deep learning-based methods rely on an indirect strategy, i.e., first establishing 2D-3D correspondences between the coordinates in the image plane and object coordinate system, and then applying a variant of the PnP/RANSAC algorithm. However, this two-stage pipeline is not end-to-end trainable, thus is hard to be employed for many tasks requiring differentiable poses. On the other hand, methods based on direct regression are currently inferior to geometry-based methods. In this work, we perform an in-depth investigation on both direct and indirect methods, and propose a simple yet effective Geometry-guided Direct Regression Network (GDR-Net) to learn the 6D pose in an end-to-end manner from dense correspondence-based intermediate geometric representations. Extensive experiments show that our approach remarkably outperforms state-of-the-art methods on LM, LM-O and YCB-V datasets. Code is available at https://git.io/GDR-Net.
****

Point Cloud Upsampling via Disentangled Refinement

Ruihui Li, Xianzhi Li, Pheng-Ann Heng, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 344-353

Point clouds produced by 3D scanning are often sparse, non-uniform, and noisy. Recent upsampling approaches aim to generate a dense point set, while achieving both distribution uniformity and proximity-to-surface, and possibly amending small holes, all in a single network. After revisiting the task, we propose to disentangle the task based on its multi-objective nature and formulate two cascaded sub-networks, a dense generator and a spatial refiner. The dense generator infers a coarse but dense output that roughly describes the underlying surface, while the spatial refiner further fine-tunes the coarse output by adjusting the location of each point. Specifically, we design a pair of local and global refinement units in the spatial refiner to evolve a coarse feature map. Also, in the spatial refiner, we regress a per-point offset vector to further adjust the coarse outputs in fine scale. Extensive qualitative and quantitative results on both synthetic and real-scanned datasets demonstrate the superiority of our method over the state-of-the-arts.
****

Feature-Level Collaboration: Joint Unsupervised Learning of Optical Flow, Stereo Depth and Camera Motion

Cheng Chi, Qingjie Wang, Tianyu Hao, Peng Guo, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2463-2473

Precise estimation of optical flow, stereo depth and camera motion are important for the real-world 3D scene understanding and visual perception. Since the three tasks are tightly coupled with the inherent 3D geometric constraints, current studies have demonstrated that the three tasks can be improved through jointly optimizing geometric loss functions of several individual networks. In this paper

, we show that effective feature-level collaboration of the networks for the three respective tasks could achieve much greater performance improvement for all three tasks than only loss-level joint optimization. Specifically, we propose a single network to combine and improve the three tasks. The network extracts the features of two consecutive stereo images, and simultaneously estimates optical flow, stereo depth and camera motion. The whole network mainly contains four parts: (I) a feature-sharing encoder to extract features of input images, which can enhance features' representation ability; (II) a pooled decoder to estimate both optical flow and stereo depth; (III) a camera pose estimation module which fuses optical flow and stereo depth information; (IV) a cost volume complement module to improve the performance of optical flow in static and occluded regions. Our method achieves state-of-the-art performance among the joint unsupervised methods, including optical flow and stereo depth estimation on KITTI 2012 and 2015 benchmarks, and camera motion estimation on KITTI VO dataset.
****
A Generalized Loss Function for Crowd Counting and Localization
Jia Wan, Ziquan Liu, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1974-1983
Previous work shows that a better density map representation can improve the performance of crowd counting. In this paper, we investigate learning the density map representation through an unbalanced optimal transport problem, and propose a generalized loss function to learn density maps for crowd counting and localization. We prove that pixel-wise L2 loss and Bayesian loss are special cases and suboptimal solutions to our proposed loss function. A perspective-guided transport cost function is further proposed to better handle the perspective transformation in crowd images. Since the predicted density will be pushed toward annotation positions, the density map prediction will be sparse and can naturally be used for localization. Finally, the proposed loss outperforms other losses on four large-scale datasets for counting, and achieves the best localization performance on NWPU-Crowd and UCF-QNRF.
****
Learning Fine-Grained Segmentation of 3D Shapes Without Part Labels
Xiaogang Wang, Xun Sun, Xinyu Cao, Kai Xu, Bin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10276-10285
Existing learning-based approaches to 3D shape segmentation usually formulate it as a semantic labeling problem, assuming that all parts of training shapes are annotated with a given set of labels. This assumption, however, is unrealistic for training fine-grained segmentation on large datasets since the annotation of fine-grained parts is extremely tedious. In this paper, we approach the problem with deep clustering, where the key idea is to learn part priors from a dataset with fine-grained segmentation but no part annotations. Given point sampled 3D shapes, we model the clustering priors of points with a similarity matrix and achieve part-based segmentation through minimizing a novel low rank loss. Further, since fine-grained parts can be very tiny, a 3D shape has to be densely sampled to ensure the tiny parts are well captured and segmented. To handle densely sampled point sets, we adopt a divide-and-conquer scheme. We first partition the large point set into a number of blocks. Each block is segmented using a deep-clustering-based part prior network (PriorNet) trained in a category-agnostic manner. We then train MergeNet, a graph convolution network, to merge the segments of all blocks to form the final segmentation result. Our method is evaluated with a challenging benchmark of fine-grained segmentation, showing significant advantage over the state-of-the-art ones.
****
Fine-Grained Shape-Appearance Mutual Learning for Cloth-Changing Person Re-Identification
Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10513-10522
Recently, person re-identification (Re-ID) has achieved great progress. However,

current methods largely depend on color appearance, which is not reliable when a person changes the clothes. Cloth-changing Re-ID is challenging since pedestrian images with clothes change exhibit large intra-class variation and small inter-class variation. Some significant features for identification are embedded in unobvious body shape differences across pedestrians. To explore such body shape cues for cloth-changing Re-ID, we propose a Fine-grained Shape-Appearance Mutual learning framework (FSAM), a two-stream framework that learns fine-grained discriminative body shape knowledge in a shape stream and transfers it to an appearance stream to complement the cloth-unrelated knowledge in the appearance features. Specifically, in the shape stream, FSAM learns fine-grained discriminative mask with the guidance of identities and extracts fine-grained body shape features by a pose-specific multi-branch network. To complement cloth-unrelated shape knowledge in the appearance stream, dense interactive mutual learning is performed across low-level and high-level features to transfer knowledge from shape stream to appearance stream, which enables the appearance stream to be deployed independently without extra computation for mask estimation. We evaluated our method on benchmark cloth-changing Re-ID datasets and achieved the start-of-the-art performance.
****
DeepSurfels: Learning Online Appearance Fusion
Marko Mihajlovic, Silvan Weder, Marc Pollefeys, Martin R. Oswald; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14524-14535
We present DeepSurfels, a novel hybrid scene representation for geometry and appearance information. DeepSurfels combines explicit and neural building blocks to jointly encode geometry and appearance information. In contrast to established representations, DeepSurfels better represents high-frequency textures, is well-suited for online updates of appearance information, and can be easily combined with machine learning methods. We further present an end-to-end trainable online appearance fusion pipeline that fuses information from RGB images into the proposed scene representation and is trained using self-supervision imposed by the reprojection error with respect to the input images. Our method compares favorably to classical texture mapping approaches as well as recent learning-based techniques. Moreover, we demonstrate lower runtime, improved generalization capabilities, and better scalability to larger scenes compared to existing methods.
****
Joint Negative and Positive Learning for Noisy Labels
Youngdong Kim, Juseung Yun, Hyounguk Shon, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9442-9451
Training of Convolutional Neural Networks (CNNs) with data with noisy labels is known to be a challenge. Based on the fact that directly providing the label to the data (Positive Learning; PL) has a risk of allowing CNNs to memorize the contaminated labels for the case of noisy data, the indirect learning approach that uses complementary labels (Negative Learning for Noisy Labels; NLNL) has proven to be highly effective in preventing overfitting to noisy data as it reduces the risk of providing faulty target. NLNL further employs a three-stage pipeline to improve convergence. As a result, filtering noisy data through the NLNL pipeline is cumbersome, increasing the training cost. In this study, we propose a novel improvement of NLNL, named Joint Negative and Positive Learning (JNPL), that unifies the filtering pipeline into a single stage. JNPL trains CNN via two losses, NL+ and PL+, which are improved upon NL and PL loss functions, respectively. We analyze the fundamental issue of NL loss function and develop new NL+ loss function producing gradient that enhances the convergence of noisy data. Furthermore, PL+ loss function is designed to enable faster convergence to expected-to-be-clean data. We show that the NL+ and PL+ train CNN simultaneously, significantly simplifying the pipeline, allowing greater ease of practical use compared to NLNL. With a simple semi-supervised training technique, our method achieves state-of-the-art accuracy for noisy data classification based on the superior filtering ability.

****
Generalizing Face Forgery Detection With High-Frequency Features
Yuchen Luo, Yong Zhang, Junchi Yan, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16317-16326
Current face forgery detection methods achieve high accuracy under the within-database scenario where training and testing forgeries are synthesized by the same algorithm. However, few of them gain satisfying performance under the cross-database scenario where training and testing forgeries are synthesized by different algorithms. In this paper, we find that current CNN-based detectors tend to overfit to method-specific color textures and thus fail to generalize. Observing that image noises remove color textures and expose discrepancies between authentic and tampered regions, we propose to utilize the high-frequency noises for face forgery detection. We carefully devise three functional modules to take full advantage of the high-frequency features. The first is the multi-scale high-frequency feature extraction module that extracts high-frequency noises at multiple scales and composes a novel modality. The second is the residual-guided spatial attention module that guides the low-level RGB feature extractor to concentrate more on forgery traces from a new perspective. The last is the cross-modality attention module that leverages the correlation between the two complementary modalities to promote feature learning for each other. Comprehensive evaluations on several benchmark databases corroborate the superior generalization performance of our proposed method.
****
The Heterogeneity Hypothesis: Finding Layer-Wise Differentiated Network Architectures
Yawei Li, Wen Li, Martin Danelljan, Kai Zhang, Shuhang Gu, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2144-2153
In this paper, we tackle the problem of convolutional neural network design. Instead of focusing on the design of the overall architecture, we investigate a design space that is usually overlooked, i.e. adjusting the channel configurations of predefined networks. We find that this adjustment can be achieved by shrinking widened baseline networks and leads to superior performance. Based on that, we articulate the "heterogeneity hypothesis": with the same training protocol, there exists a layer-wise differentiated network architecture (LW-DNA) that can outperform the original network with regular channel configurations but with a lower level of model complexity. The LW-DNA models are identified without extra computational cost or training time compared with the original network. This constraint leads to controlled experiments which direct the focus to the importance of layer-wise specific channel configurations. LW-DNA models come with advantages related to overfitting, i.e. the relative relationship between model complexity and dataset size. Experiments are conducted on various networks and datasets for image classification, visual tracking and image restoration. The resultant LW-DNA models consistently outperform the baseline models. Code is available at https://github.com/ofsoundof/Heterogeneity_Hypothesis.git.
****
Robust Neural Routing Through Space Partitions for Camera Relocalization in Dynamic Indoor Environments
Siyan Dong, Qingnan Fan, He Wang, Ji Shi, Li Yi, Thomas Funkhouser, Baoquan Chen, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8544-8554
Localizing the camera in a known indoor environment is a key building block for scene mapping, robot navigation, AR, etc. Recent advances estimate the camera pose via optimization over the 2D/3D-3D correspondences established between the coordinates in 2D/3D camera space and 3D world space. Such a mapping is estimated with either a convolution neural network or a decision tree using only the static input image sequence, which makes these approaches vulnerable to dynamic indoor environments that are quite common yet challenging in the real world. To address the aforementioned issues, in this paper, we propose a novel outlier-aware neural tree which bridges the two worlds, deep learning and decision tree approach

es. It builds on three important blocks: (a) a hierarchical space partition over the indoor scene to construct the decision tree; (b) a neural routing function, implemented as a deep classification network, employed for better 3D scene unde rstanding; and (c) an outlier rejection module used to filter out dynamic points during the hierarchical routing process. Our proposed algorithm is evaluated on the RIO-10 benchmark developed for camera relocalization in dynamic indoor envi ronments. It achieves robust neural routing through space partitions and outperf orms the state-of-the-art approaches by around 30% on camera pose accuracy, whil e running comparably fast for evaluation.
****
Facial Action Unit Detection With Transformers
Geethu Miriam Jacob, Bjorn Stenger; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2021, pp. 7680-7689
The Facial Action Coding System is a taxonomy for fine-grained facial expression analysis. This paper proposes a method for detecting Facial Action Units (FAU), which define particular face muscle activity, from an input image. FAU detectio n is formulated as a multi-task learning problem, where image features and atten tion maps are input to a branch for each action unit to extract discriminative f eature embeddings, using a new loss function, the Center Contrastive (CC) loss. We employ a new FAU correlation network, based on a transformer encoder architec ture, to capture the relationships between different action units for the wide r ange of expressions in the training data. The resulting features are shown to yi eld high classification performance. We validate our design choices, including t he use of CC loss and Tversky loss functions, in ablative experiments. We show t hat the proposed method outperforms state-of-theart techniques on two public dat asets, BP4D and DISFA, with an absolute improvement of the F1-score of over 2% o n each.
****
Exploiting Aliasing for Manga Restoration
Minshan Xie, Menghan Xia, Tien-Tsin Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13405-13414
As a popular entertainment art form, manga enriches the line drawings details wi th bitonal screentones. However, manga resources over the Internet usually show screentone artifacts because of inappropriate scanning/rescaling resolution. In this paper, we propose an innovative two-stage method to restore quality bitonal manga from degraded ones. Our key observation is that the aliasing induced by d ownsampling bitonal screentones can be utilized as informative clues to infer th e original resolution and screentones. First, we predict the target resolution f rom the degraded manga via the Scale Estimation Network (SE-Net) with spatial vo ting scheme. Then, at the target resolution, we restore the region-wise bitonal screentones via the Manga Restoration Network (MR-Net) discriminatively, dependi ng on the degradation degree. Specifically, the original screentones are directl y restored in pattern-identifiable regions, and visually plausible screentones a re synthesized in pattern-agnostic regions. Quantitative evaluation on synthetic data and visual assessment on real-world cases illustrate the effectiveness of our method.
****
Discovering Hidden Physics Behind Transport Dynamics
Peirong Liu, Lin Tian, Yubo Zhang, Stephen Aylward, Yueh Lee, Marc Niethammer; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10082-10092
Transport processes are ubiquitous. They are, for example, at the heart of optic al flow approaches; or of perfusion imaging, where blood transport is assessed, most commonly by injecting a tracer. An advection-diffusion equation is widely u sed to describe these transport phenomena. Our goal is estimating the underlying physics of advection-diffusion equations, expressed as velocity and diffusion t ensor fields. We propose a learning framework (YETI) building on an auto-encoder structure between 2D and 3D image time-series, which incorporates the advection -diffusion model. To help with identifiability, we develop an advection-diffusio n simulator which allows pre-training of our model by supervised learning using

the velocity and diffusion tensor fields. Instead of directly learning these velocity and diffusion tensor fields, we introduce representations that assure incompressible flow and symmetric positive semi-definite diffusion fields and demonstrate the additional benefits of these representations on improving estimation accuracy. We further use transfer learning to apply YETI on a public brain magnetic resonance (MR) perfusion dataset of stroke patients and show its ability to successfully distinguish stroke lesions from normal brain regions via the estimated velocity and diffusion tensor fields.
****
Cross-View Gait Recognition With Deep Universal Linear Embeddings
Shaoxiong Zhang, Yunhong Wang, Annan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9095-9104
Gait is considered an attractive biometric identifier for its non-invasive and non-cooperative features compared with other biometric identifiers such as fingerprint and iris. At present, cross-view gait recognition methods always establish representations from various deep convolutional networks for recognition and ignore the potential dynamical information of the gait sequences. If assuming that pedestrians have different walking patterns, gait recognition can be performed by calculating their dynamical features from each view. This paper introduces the Koopman operator theory to gait recognition, which can find an embedding space for a global linear approximation of a nonlinear dynamical system. Furthermore, a novel framework based on convolutional variational autoencoder and deep Koopman embedding is proposed to approximate the Koopman operators, which is used as dynamical features from the linearized embedding space for cross-view gait recognition. It gives solid physical interpretability for a gait recognition system. Experiments on a large public dataset, OU-MVLP, prove the effectiveness of the proposed method.
****
Tuning IR-Cut Filter for Illumination-Aware Spectral Reconstruction From RGB
Bo Sun, Junchi Yan, Xiao Zhou, Yinqiang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 84-93
To reconstruct spectral signals from multi-channel observations, in particular trichromatic RGBs, has recently emerged as a promising alternative to traditional scanning-based spectral imager. It has been proven that the reconstruction accuracy relies heavily on the spectral response of the RGB camera in use. To improve accuracy, data-driven algorithms have been proposed to retrieve the best response curves of existing RGB cameras, or even to design brand new three-channel response curves. Instead, this paper explores the filter-array based color imaging mechanism of existing RGB cameras, and proposes to design the IR-cut filter properly for improved spectral recovery, which stands out as an in-between solution with better trade-off between reconstruction accuracy and implementation complexity. We further propose a deep learning based spectral reconstruction method, which allows to recover the illumination spectrum as well. Experiment results with both synthetic and real images under daylight illumination have shown the benefits of our IR-cut filter tuning method and our illumination-aware spectral reconstruction method.
****
Relative Order Analysis and Optimization for Unsupervised Deep Metric Learning
Shichao Kan, Yigang Cen, Yang Li, Vladimir Mladenovic, Zhihai He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13999-14008
In unsupervised learning of image features without labels, especially on datasets with fine-grained object classes, it is often very difficult to tell if a given image belongs to one specific object class or another, even for human eyes. However, we can reliably tell if image C is more similar to image A than image B. In this work, we propose to explore how this relative order can be used to learn discriminative features with an unsupervised metric learning method. Instead of resorting to clustering or self-supervision to create pseudo labels for an absolute decision, which often suffers from high label error rates, we construct reliable relative orders for groups of image samples and learn a deep neural networ

k to predict these relative orders. During training, this relative order prediction network and the feature embedding network are tightly coupled, providing mutual constraints to each other to improve overall metric learning performance in a cooperative manner. During testing, the predicted relative orders are used as constraints to optimize the generated features and refine their feature distance-based image retrieval results using a constrained optimization procedure. Our experimental results demonstrate that the proposed relative orders for unsupervised learning (ROUL) method is able to significantly improve the performance of unsupervised deep metric learning.
****
Anchor-Free Person Search
Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7690-7699
Person search aims to simultaneously localize and identify a query person from realistic, uncropped images, which can be regarded as the unified task of pedestrian detection and person re-identification (re-id). Most existing works employ two-stage detectors like Faster-RCNN, yielding encouraging accuracy but with high computational overhead. In this work, we present the Feature-Aligned Person Search Network (AlignPS), the first anchor-free framework to efficiently tackle this challenging task. AlignPS explicitly addresses the major challenges, which we summarize as the misalignment issues in different levels (i.e., scale, region, and task), when accommodating an anchor-free detector for this task. More specifically, we propose an aligned feature aggregation module to generate more discriminative and robust feature embeddings by following a "re-id first" principle. Such a simple design directly improves the baseline anchor-free model on CUHK-SYSU by more than 20% in mAP. Moreover, AlignPS outperforms state-of-the-art two-stage methods, with a higher speed. The code is available at https://github.com/daodaofr/AlignPS.
****
Are Labels Always Necessary for Classifier Accuracy Evaluation?
Weijian Deng, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15069-15078
To calculate the model accuracy on a computer vision task, e.g., object recognition, we usually require a test set composing of test samples and their ground truth labels. Whilst standard usage cases satisfy this requirement, many real-world scenarios involve unlabeled test data, rendering common model evaluation methods infeasible. We investigate this important and under-explored problem, Automatic model Evaluation (AutoEval). Specifically, given a labeled training set and a classifier, we aim to estimate the classification accuracy on unlabeled test datasets. We construct a meta-dataset: a dataset comprised of datasets generated from the original images via various transformations such as rotation, background substitution, foreground scaling, etc. As the classification accuracy of the model on each sample (dataset) is known from the original dataset labels, our task can be solved via regression. Using the feature statistics to represent the distribution of a sample dataset, we can train regression models (e.g., a regression neural network) to predict model performance. Using synthetic meta-dataset and real-world datasets in training and testing, respectively, we report a reasonable and promising prediction of the model accuracy. We also provide insights into the application scope, limitation, and potential future direction of AutoEval.
****
Self-Supervised Motion Learning From Static Images
Ziyuan Huang, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Rong Jin, Marcelo H. Ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1276-1285
Motions are reflected in videos as the movement of pixels, and actions are essentially patterns of inconsistent motions between the foreground and the background. To well distinguish the actions, especially those with complicated spatio-temporal interactions, correctly locating the prominent motion areas is of crucial importance. However, most motion information in existing videos are difficult to

label and training a model with good motion representations with supervision wi ll thus require a large amount of human labour for annotation. In this paper, we address this problem by self-supervised learning. Specifically, we propose to l earn Motion from Static Images (MoSI). The model learns to encode motion informa tion by classifying pseudo motions generated by MoSI. We furthermore introduce a static mask in pseudo motions to create local motion patterns, which forces the model to additionally locate notable motion areas for the correct classificatio n.We demonstrate that MoSI can discover regions with large motion even without f ine-tuning on the downstream datasets. As a result, the learned motion represent ations boost the performance of tasks requiring understanding of complex scenes and motions, i.e., action recognition. Extensive experiments show the consistent and transferable improvements achieved by MoSI. Codes will be soon released.
****
AttentiveNAS: Improving Neural Architecture Search via Attentive Sampling
Dilin Wang, Meng Li, Chengyue Gong, Vikas Chandra; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6418-6427
Neural architecture search (NAS) has shown great promise in designing state-of-t he-art (SOTA) models that are both accurate and efficient. Recently, two-stage N AS, e.g. BigNAS, decouples the model training and searching process and achieves remarkable search efficiency and accuracy. Two-stage NAS requires sampling from the search space during training, which directly impacts the accuracy of the fi nal searched models. While uniform sampling has been widely used for its simplic ity, it is agnostic of the model performance Pareto front, which is the main foc us in the search process, and thus, misses opportunities to further improve the model accuracy. In this work, we propose AttentiveNAS that focuses on improving the sampling strategy to achieve better performance Pareto. We also propose algo rithms to efficiently and effectively identify the networks on the Pareto during training. Without extra re-training or post-processing, we can simultaneously o btain a large number of networks across a wide range of FLOPs. Our discovered mo del family, AttentiveNAS models, achieves top-1 accuracy from 77.3% to 80.7% on ImageNet, and outperforms SOTA models, including BigNAS, Once-for-All networks a nd FBNetV3. We also achieve ImageNet accuracy of 80.1% with only 491 MFLOPs. Our training code and pretrained models are available at https://github. com/facebo okresearch/AttentiveNAS.
****
StablePose: Learning 6D Object Poses From Geometrically Stable Patches
Yifei Shi, Junwen Huang, Xin Xu, Yifan Zhang, Kai Xu; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15222-15231
We introduce the concept of geometric stability to the problem of 6D object pose estimation and propose to learn pose inference based on geometrically stable pa tches extracted from observed 3D point clouds. According to the theory of geomet ric stability analysis, a minimal set of three planar/cylindrical patches are ge ometrically stable and determine the full 6DoFs of the object pose. We train a d eep neural network to regress 6D object pose based on geometrically stable patch groups via learning both intra-patch geometric features and inter-patch context ual features. A subnetwork is jointly trained to predict per-patch poses. This a uxiliary task is a relaxation of the group pose prediction: A single patch canno t determine the full 6DoFs but is able to improve pose accuracy in its correspon ding DoFs. Working with patch groups makes our method generalize well for random occlusion and unseen instances. The method is easily amenable to resolve symmet ry ambiguities. Our method achieves the state-of-the-art results on public bench marks compared not only to depth-only but also to RGBD methods. It also performs well in category-level pose estimation.
****
Towards Evaluating and Training Verifiably Robust Neural Networks
Zhaoyang Lyu, Minghao Guo, Tong Wu, Guodong Xu, Kehuan Zhang, Dahua Lin; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4308-4317
Recent works have shown that interval bound propagation (IBP) can be used to tra

in verifiably robust neural networks. Reseachers observe an intriguing phenomeno
n on these IBP trained networks: CROWN, a bounding method based on tight linear
relaxation, often gives very loose bounds on these networks. We also observe tha
t most neurons become dead during the IBP training process, which could hurt the
 representation capability of the network. In this paper, we study the relations
hip between IBP and CROWN, and prove that CROWN is always tighter than IBP when
choosing appropriate bounding lines. We further propose a relaxed version of CRO
WN, linear bound propagation (LBP), that can be used to verify large networks to
 obtain lower verified errors than IBP. We also design a new activation function
, parameterized ramp function (ParamRamp), which has more diversity of neuron st
atus than ReLU. We conduct extensive experiments on MNIST, CIFAR-10 and Tiny-Ima
geNet with ParamRamp activation and achieve state-of-the-art verified robustness
. Code is available at https://github.com/ZhaoyangLyu/VerifiablyRobustNN.
****
Interpolation-Based Semi-Supervised Learning for Object Detection
Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, Nojun Kwak; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021
, pp. 11602-11611
Despite the data labeling cost for the object detection tasks being substantiall
y more than that of the classification tasks, semi-supervised learning methods f
or object detection have not been studied much. In this paper, we propose an Int
erpolation-based Semi-supervised learning method for object Detection (ISD), whi
ch considers and solves the problems caused by applying conventional Interpolati
on Regularization (IR) directly to object detection. We divide the output of the
 model into two types according to the objectness scores of both original patche
s that are mixed in IR. Then, we apply a separate loss suitable for each type in
 an unsupervised manner. The proposed losses dramatically improve the performanc
e of semi-supervised learning as well as supervised learning. In the supervised
learning setting, our method improves the baseline methods by a significant marg
in. In the semi-supervised learning setting, our algorithm improves the performa
nce on a benchmark dataset (PASCAL VOC and MSCOCO) in a benchmark architecture (
SSD).
****
Teachers Do More Than Teach: Compressing Image-to-Image Models
Qing Jin, Jian Ren, Oliver J. Woodford, Jiazhuo Wang, Geng Yuan, Yanzhi Wang, Se
rgey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2021, pp. 13600-13611
Generative Adversarial Networks (GANs) have achieved huge success in generating
high-fidelity images, however, they suffer from low efficiency due to tremendous
 computational cost and bulky memory usage. Recent efforts on compression GANs s
how noticeable progress in obtaining smaller generators by sacrificing image qua
lity or involving a time-consuming searching process. In this work, we aim to ad
dress these issues by introducing a teacher network that provides a search space
 in which efficient network architectures can be found, in addition to performin
g knowledge distillation. First, we revisit the search space of generative model
s, introducing an inception-based residual block into generators. Second, to ach
ieve target computation cost, we propose a one-step pruning algorithm that searc
hes a student architecture from the teacher model and substantially reduces sear
ching cost. It requires no L1 sparsity regularization and its associated hyper-p
arameters, simplifying the training procedure. Finally, we propose to distill kn
owledge through maximizing feature similarity between teacher and student via an
 index named Global Centered Kernel Alignment (GCKA). Our compressed networks ac
hieve better image fidelity (FID, mIoU) than the original models with much-reduc
ed computational cost, e.g., MACs.
****
Seeing in Extra Darkness Using a Deep-Red Flash
Jinhui Xiong, Jian Wang, Wolfgang Heidrich, Shree Nayar; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 100
00-10009
We propose a new flash technique for low-light imaging, using deep-red light as

an illuminating source. Our main observation is that in a dim environment, the h
uman eye mainly uses rods for the perception of light, which are not sensitive t
o wavelengths longer than 620nm, yet the camera sensor still has a spectral resp
onse. We propose a novel modulation strategy when training a modern CNN model fo
r guided image filtering, fusing a noisy RGB frame and a flash frame. This fusio
n network is further extended for video reconstruction. We have built a prototyp
e with minor hardware adjustments and tested the new flash technique on a variet
y of static and dynamic scenes. The experimental results demonstrate that our me
thod produces compelling reconstructions, even in extra dim conditions.
****
PSD: Principled Synthetic-to-Real Dehazing Guided by Physical Priors
Zeyuan Chen, Yangchao Wang, Yang Yang, Dong Liu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7180-7189
Deep learning-based methods have achieved remarkable performance for image dehaz
ing. However, previous studies are mostly focused on training models with synthe
tic hazy images, which incurs performance drop when the models are used for real
-world hazy images. We propose a Principled Synthetic-to-real Dehazing (PSD) fra
mework to improve the generalization performance of dehazing. Starting from a de
hazing model backbone that is pre-trained on synthetic data, PSD exploits real h
azy images to fine-tune the model in an unsupervised fashion. For the fine-tunin
g, we leverage several well-grounded physical priors and combine them into a pri
or loss committee. PSD allows for most of the existing dehazing models as its ba
ckbone, and the combination of multiple physical priors boosts dehazing signific
antly. Through extensive experiments, we demonstrate that our PSD framework esta
blishes the new state-of-the-art performance for real-world dehazing, in terms o
f visual quality assessed by no-reference quality metrics as well as subjective
evaluation and downstream task performance indicator.
****
3D Spatial Recognition Without Spatially Labeled 3D
Zhongzheng Ren, Ishan Misra, Alexander G. Schwing, Rohit Girdhar; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021
, pp. 13204-13213
We introduce WyPR, a Weakly-supervised framework for Point cloud Recognition, re
quiring only scene-level class tags as supervision. WyPR jointly addresses three
 core 3D recognition tasks: point-level semantic segmentation, 3D proposal gener
ation, and 3D object detection, coupling their predictions through self and cros
s-task consistency losses. We show that in conjunction with standard multiple-in
stance learning objectives, WyPR can detect and segment objects in point cloud w
ithout access to any spatial labels at training time. We demonstrate its efficac
y using the ScanNet and S3DIS datasets, outperforming prior state of the art on
weakly-supervised segmentation by more than 6% mIoU. In addition, we set up the
first benchmark for weakly-supervised 3D object detection on both datasets, wher
e WyPR outperforms standard approaches and establishes strong baselines for futu
re work.
****
Robust Reference-Based Super-Resolution via C2-Matching
Yuming Jiang, Kelvin C.K. Chan, Xintao Wang, Chen Change Loy, Ziwei Liu; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2021, pp. 2103-2112
Reference-based Super-Resolution (Ref-SR) has recently emerged as a promising pa
radigm to enhance a low-resolution (LR) input image by introducing an additional
 high-resolution (HR) reference image. Existing Ref-SR methods mostly rely on im
plicit correspondence matching to borrow HR textures from reference images to co
mpensate for the information loss in input images. However, performing local tra
nsfer is difficult because of two gaps between input and reference images: the t
ransformation gap (e.g. scale and rotation) and the resolution gap (e.g. HR and
LR). To tackle these challenges, we propose C^ 2 -Matching in this work, which p
roduces explicit robust matching crossing transformation and resolution. 1) For
the transformation gap, we propose a contrastive correspondence network, which l
earns transformation-robust correspondences using augmented views of the input i

mage. 2) For the resolution gap, we adopt a teacher-student correlation distillation, which distills knowledge from the easier HR-HR matching to guide the more ambiguous LR-HR matching. 3) Finally, we design a dynamic aggregation module to address the potential misalignment issue. In addition, to faithfully evaluate the performance of Ref-SR under a realistic setting, we contribute the Webly-Referenced SR (WR-SR) dataset, mimicking the practical usage scenario. Extensive experiments demonstrate that our proposed C^ 2 -Matching significantly outperforms current state-of-the-art methods by over 1dB on the standard CUFED5 benchmark. Notably, it also shows great generalizability on WR-SR dataset as well as robustness across large scale and rotation transformations.
****
Temporal-Relational CrossTransformers for Few-Shot Action Recognition
Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 475-484
We propose a novel approach to few-shot action recognition, finding temporally-corresponding frame tuples between the query and videos in the support set. Distinct from previous few-shot works, we construct class prototypes using the CrossTransformer attention mechanism to observe relevant sub-sequences of all support videos, rather than using class averages or single best matches. Video representations are formed from ordered tuples of varying numbers of frames, which allows sub-sequences of actions at different speeds and temporal offsets to be compared. Our proposed Temporal-Relational CrossTransformers (TRX) achieve state-of-the-art results on few-shot splits of Kinetics, Something-Something V2 (SSv2), HMDB51 and UCF101. Importantly, our method outperforms prior work on SSv2 by a wide margin (12%) due to the its ability to model temporal relations. A detailed ablation showcases the importance of matching to multiple support set videos and learning higher-order relational CrossTransformers.
****
Understanding Failures of Deep Networks via Robust Feature Extraction
Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, Eric Horvitz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12853-12862
Traditional evaluation metrics for learned models that report aggregate scores over a test set are insufficient for surfacing important and informative patterns of failure over features and instances. We introduce and study a method aimed at characterizing and explaining failures by identifying visual attributes whose presence or absence results in poor performance. In distinction to previous work that relies upon crowdsourced labels for visual attributes, we leverage the representation of a separate robust model to extract interpretable features and then harness these features to identify failure modes. We further propose a visualization method aimed at enabling humans to understand the meaning encoded in such features and we test the comprehensibility of the features. An evaluation of the methods on the ImageNet dataset demonstrates that: (i) the proposed workflow is effective for discovering important failure modes, (ii) the visualization techniques help humans to understand the extracted features, and (iii) the extracted insights can assist engineers with error analysis and debugging.
****
Relation-aware Instance Refinement for Weakly Supervised Visual Grounding
Yongfei Liu, Bo Wan, Lin Ma, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5612-5621
Visual grounding, which aims to build a correspondence between visual objects and their language entities, plays a key role in cross-modal scene understanding. One promising and scalable strategy for learning visual grounding is to utilize weak supervision from only image-caption pairs. Previous methods typically rely on matching query phrases directly to a precomputed, fixed object candidate pool, which leads to inaccurate localization and ambiguous matching due to lack of semantic relation constraints. In our paper, we propose a novel context-aware weakly-supervised learning method that incorporates coarse-to-fine object refinement and entity relation modeling into a two-stage deep network, capable of produci

ng more accurate object representation and matching. To effectively train our network, we introduce a self-taught regression loss for the proposal locations and a classification loss based on parsed entity relations. Extensive experiments on two public benchmarks Flickr30K Entities and ReferItGame demonstrate the efficacy of our weakly grounding framework. The results show that we outperform the previous methods by a considerable margin, achieving 59.27% top-1 accuracy in Flickr30K Entities and 37.68% in the ReferItGame dataset respectively.
****

Spatially-Invariant Style-Codes Controlled Makeup Transfer
Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6549-6557
Transferring makeup from the misaligned reference image is challenging. Previous methods overcome this barrier by computing pixel-wise correspondences between two images, which is inaccurate and computational-expensive. In this paper, we take a different perspective to break down the makeup transfer problem into a two-step extraction-assignment process. To this end, we propose a Style-based Controllable GAN model that consists of three components, each of which corresponds to target style-code encoding, face identity features extraction, and makeup fusion, respectively. In particular, a Part-specific Style Encoder encodes the component-wise makeup style of the reference image into a style-code in an intermediate latent space W. The style-code discards spatial information and therefore is invariant to spatial misalignment. On the other hand, the style-code embeds component-wise information, enabling flexible partial makeup editing from multiple references. This style-code, together with source identity features, are integrated to a Makeup Fusion Decoder equipped with multiple AdaIN layers to generate the final result. Our proposed method demonstrates great flexibility on makeup transfer by supporting makeup removal, shade-controllable makeup transfer, and part-specific makeup transfer, even with large spatial misalignment. Extensive experiments demonstrate the superiority of our approach over state-of-the-art methods. Code is available at https://github.com/makeuptransfer/SCGAN.
****

Adaptive Image Transformer for One-Shot Object Detection
Ding-Jie Chen, He-Yen Hsieh, Tyng-Luh Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12247-12256
One-shot object detection tackles a challenging task that aims at identifying within a target image all object instances of the same class, implied by a query image patch. The main difficulty lies in the situation that the class label of the query patch and its respective examples are not available in the training data. Our main idea leverages the concept of language translation to boost metric-learning-based detection methods. Specifically, we emulate the language translation process to adaptively translate the feature of each object proposal to better correlate the given query feature for discriminating the class-similarity among the proposal-query pairs. To this end, we propose the Adaptive Image Transformer (AIT) module that deploys an attention-based encoder-decoder architecture to simultaneously explore intra-coder and inter-coder (i.e., each proposal-query pair) attention. The adaptive nature of our design turns out to be flexible and effective in addressing the one-shot learning scenario. With the informative attention cues, the proposed model excels in predicting the class-similarity between the target image proposals and the query image patch. Though conceptually simple, our model significantly outperforms a state-of-the-art technique, improving the unseen-class object classification from 63.8 mAP and 22.0 AP50 to 72.2 mAP and 24.3 AP50 on the PASCAL-VOC and MS-COCO benchmark datasets, respectively.
****

Bilateral Grid Learning for Stereo Matching Networks
Bin Xu, Yuhua Xu, Xiaoli Yang, Wei Jia, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12497-12506
Real-time performance of stereo matching networks is important for many applications, such as automatic driving, robot navigation and augmented reality (AR). Al

though significant progress has been made in stereo matching networks in recent
years, it is still challenging to balance real-time performance and accuracy. In
 this paper, we present a novel edge-preserving cost volume upsampling module ba
sed on the slicing operation in the learned bilateral grid. The slicing layer is
 parameter-free, which allows us to obtain a high quality cost volume of high re
solution from a low-resolution cost volume under the guide of the learned guidan
ce map efficiently. The proposed cost volume upsampling module can be seamlessly
 embedded into many existing stereo matching networks, such as GCNet, PSMNet, an
d GANet. The resulting networks are accelerated several times while maintaining
comparable accuracy. Furthermore, we design a real-time network (named BGNet) ba
sed on this module, which outperforms existing published real-time deep stereo m
atching networks, as well as some complex networks on the KITTI stereo datasets.
 The code is available at https://github.com/YuhuaXu/BGNet.
****
A Multi-Task Network for Joint Specular Highlight Detection and Removal
Gang Fu, Qing Zhang, Lei Zhu, Ping Li, Chunxia Xiao; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7752-77
61
Specular highlight detection and removal are fundamental and challenging tasks.
Although recent methods achieve promising results on the two tasks by supervised
 training on synthetic training data, they are typically solely designed for hig
hlight detection or removal, and their performance usually deteriorates signific
antly on real-world images. In this paper, we present a novel network that aims
to detect and remove highlights from natural images. To remove the domain gap be
tween synthetic training samples and real test images, and support the investiga
tion of learning-based approaches, we first introduce a dataset of 16K real imag
es, each of which has the corresponding highlight detection and removal images.
Using the presented dataset, we develop a multi-task network for joint highlight
 detection and removal, based on a new specular highlight image formation model.
 Experiments on the benchmark datasets and our new dataset show that our approac
h clearly outperforms the state-of-the-art methods for both highlight detection
and removal.
****
A Deep Emulator for Secondary Motion of 3D Characters
Mianlun Zheng, Yi Zhou, Duygu Ceylan, Jernej Barbic; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5932-59
40
Fast and light-weight methods for animating 3D characters are desirable in vario
us applications such as computer games. We present a learning-based approach to
enhance skinning-based animations of 3D characters with vivid secondary motion e
ffects. We represent each local patch of a character simulation mesh as a graph
network where the edges implicitly encode the internal forces between the neighb
oring vertices. We then train a neural network that emulates the ordinary differ
ential equations of the character dynamics, predicting new vertex positions from
 the current accelerations, velocities and positions. Being a local method, our
network is independent of the mesh topology and generalizes to arbitrarily shape
d 3D character meshes at test time. We further represent per-vertex constraints
and material properties such as stiffness, enabling us to easily adjust the dyna
mics in different parts of the mesh. We evaluate our method on various character
 meshes and complex motion sequences. Our method can be over 30 times more effic
ient than ground-truth physically based simulation, and outperforms alternative
solutions that provide fast approximations.
****
Omni-Supervised Point Cloud Segmentation via Gradual Receptive Field Component R
easoning
Jingyu Gong, Jiachen Xu, Xin Tan, Haichuan Song, Yanyun Qu, Yuan Xie, Lizhuang M
a; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2021, pp. 11673-11682
Hidden features in neural network usually fail to learn informative representati
on for 3D segmentation as supervisions are only given on output prediction, whil

e this can be solved by omni-scale supervision on intermediate layers. In this paper, we bring the first omni-scale supervision method to point cloud segmentation via the proposed gradual Receptive Field Component Reasoning (RFCR), where target Receptive Field Component Codes (RFCCs) are designed to record categories within receptive fields for hidden units in the encoder. Then, target RFCCs will supervise the decoder to gradually infer the RFCCs in a coarse-to-fine categories reasoning manner, and finally obtain the semantic labels. Because many hidden features are inactive with tiny magnitude and make minor contributions to RFCC prediction, we propose a Feature Densification with a centrifugal potential to obtain more unambiguous features, and it is in effect equivalent to entropy regularization over features. More active features can further unleash the potential of our omni-supervision method. We embed our method into four prevailing backbones and test on three challenging benchmarks. Our method can significantly improves the backbones in all three datasets. Specifically, our method brings new state-of-the-art performances for S3DIS as well as Semantic3D and ranks the 1st in the ScanNet benchmark among all the point-based methods. Code is publicly available at https://github.com/azuki-miho/RFCR.
****
All Labels Are Not Created Equal: Enhancing Semi-Supervision via Label Grouping and Co-Training
Islam Nassar, Samitha Herath, Ehsan Abbasnejad, Wray Buntine, Gholamreza Haffari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7241-7250
Pseudo-labeling is a key component in semi-supervised learning (SSL). It relies on iteratively using the model to generate artificial labels for the unlabeled data to train against. A common property among its various methods is that they only rely on the model's prediction to make labeling decisions without considering any prior knowledge about the visual similarity among the classes. In this paper, we demonstrate that this degrades the quality of pseudo-labeling as it poorly represents visually similar classes in the pool of pseudo-labeled data. We propose SemCo, a method which leverages label semantics and co-training to address this problem. We train two classifiers with two different views of the class labels: one classifier uses the one-hot view of the labels and disregards any potential similarity among the classes, while the other uses a distributed view of the labels and groups potentially similar classes together. We then co-train the two classifiers to learn based on their disagreements. We show that our method achieves state-of-the-art performance across various SSL tasks including 5.6% accuracy improvement on Mini-ImageNet dataset with 1000 labeled examples. We also show that our method requires smaller batch size and fewer training iterations to reach its best performance. We make our code available at https://github.com/islam-nassar/semco.
****
PMP-Net: Point Cloud Completion by Learning Multi-Step Point Moving Paths
Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, Yu-Shen Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7443-7452
The task of point cloud completion aims to predict the missing part for an incomplete 3D shape. A widely used strategy is to generate a complete point cloud from the incomplete one. However, the unordered nature of point clouds will degrade the generation of high-quality 3D shapes, as the detailed topology and structure of discrete points are hard to be captured by the generative process only using a latent code. In this paper, we address the above problem by reconsidering the completion task from a new perspective, where we formulate the prediction as a point cloud deformation process. Specifically, we design a novel neural network, named PMP-Net, to mimic the behavior of an earth mover. It moves move each point of the incomplete input to complete the point cloud, where the total distance of point moving paths (PMP) should be shortest. Therefore, PMP-Net predicts a unique point moving path for each point according to the constraint of total point moving distances. As a result, the network learns a strict and unique correspondence on point-level, and thus improves the quality of the predicted complete s

hape. We conduct comprehensive experiments on Completion3D and PCN datasets, whi
ch demonstrate our advantages over the state-of-the-art point cloud completion m
ethods. Code will be available at https://github.com/diviswen/PMP-Net.
****
Gradient-Based Algorithms for Machine Teaching
Pei Wang, Kabir Nagrecha, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1387-1396
The problem of machine teaching is considered. A new formulation is proposed und
er the assumption of an optimal student, where optimality is defined in the usua
l machine learning sense of empirical risk minimization. This is a sensible assu
mption for machine learning students and for human students in crowdsourcing pla
tforms, who tend to perform at least as well as machine learning systems. It is
shown that, if allowed unbounded effort, the optimal student always learns the o
ptimal predictor for a classification task. Hence, the role of the optimal teach
er is to select the teaching set that minimizes student effort. This is formulat
ed as a problem of functional optimization where, at each teaching iteration, th
e teacher seeks to align the steepest descent directions of the risk of (1) the
teaching set and (2) entire example population. The optimal teacher, denoted Max
Grad, is then shown to maximize the gradient of the risk on the set of new examp
les selected per iteration. MaxGrad teaching algorithms are finally provided for
 both binary and multiclass tasks, and shown to have some similarities with boos
ting algorithms. Experimental evaluations demonstrate the effectiveness of MaxGr
ad, which outperforms previous algorithms on the classification task, for both m
achine learning and human students from MTurk, by a substantial margin.
****
MetaSCI: Scalable and Adaptive Reconstruction for Video Compressive Sensing
Zhengjue Wang, Hao Zhang, Ziheng Cheng, Bo Chen, Xin Yuan; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2
083-2092
To capture high-speed videos using a two-dimensional detector, video snapshot co
mpressive imaging (SCI) is a promising system, where the video frames are coded
by different masks and then compressed to a snapshot measurement. Following this
, efficient algorithms are desired to reconstruct the high-speed frames, where t
he state-of-the-art results are achieved by deep learning networks. However, the
se networks are usually trained for specific small-scale masks and often have hi
gh demands of training time and GPU memory, which are hence not flexible to i) a
 new mask with the same size and ii) a larger-scale mask. We address these chall
enges by developing a Meta Modulated Convolutional Network for SCI reconstructio
n, dubbed MetaSCI. MetaSCI is composed of a shared backbone for different masks,
 and light-weight meta-modulation parameters to evolve to different modulation p
arameters for each mask, thus having the properties of fast adaptation to new ma
sks (or systems) and ready to scale to large data. Extensive simulation and real
 data results demonstrate the superior performance of our proposed approach.
****
Removing Raindrops and Rain Streaks in One Go
Ruijie Quan, Xin Yu, Yuanzhi Liang, Yi Yang; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9147-9156
Existing rain-removal algorithms often tackle either rain streak removal or rain
drop removal, and thus may fail to handle real-world rainy scenes. Besides, the
lack of real-world deraining datasets comprising different types of rain and the
ir corresponding rain-free ground-truth also impedes deraining algorithm develop
ment. In this paper, we aim to address real-world deraining problems from two as
pects. First, we propose a complementary cascaded network architecture, namely C
CN, to remove rain streaks and raindrops in a unified framework. Specifically, o
ur CCN removes raindrops and rain streaks in a complementary fashion, i.e., rain
drop removal followed by rain streak removal and vice versa, and then fuses the
results via an attention based fusion module. Considering significant shape and
structure differences between rain streaks and raindrops, it is difficult to man
ually design a sophisticated network to remove them effectively. Thus, we employ
 neural architecture search to adaptively find optimal architectures within our

specified deraining search space. Second, we present a new real-world rain datas et, namely RainDS, to prosper the development of deraining algorithms in practic al scenarios. RainDS consists of rain images in different types and their corres ponding rain-free ground-truth, including rain streak only, raindrop only, and b oth of them. Extensive experimental results on both existing benchmarks and Rain DS demonstrate that our method outperforms the state-of-the-art.
****
Action Unit Memory Network for Weakly Supervised Temporal Action Localization
Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, Yongdong Zha ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2021, pp. 9969-9979
Weakly supervised temporal action localization aims to detect and localize actio ns in untrimmed videos with only video-level labels during training. However, wi thout frame-level annotations, it is challenging to achieve localization complet eness and relieve background interference. In this paper, we present an Action U nit Memory Network (AUMN) for weakly supervised temporal action localization, wh ich can mitigate the above two challenges by learning an action unit memory bank . In the proposed AUMN, two attention modules are designed to update the memory bank adaptively and learn action units specific classifiers. Furthermore, three effective mechanisms (diversity, homogeneity and sparsity) are designed to guide the updating of the memory network. To the best of our knowledge, this is the f irst work to explicitly model the action units with a memory network. Extensive experimental results on two standard benchmarks (THUMOS14 and ActivityNet) demon strate that our AUMN performs favorably against stateof-the-art methods. Specifi cally, the average mAP of IoU thresholds from 0.1 to 0.5 on the THUMOS14 dataset is significantly improved from 47.0% to 52.1%.
****
IMAGINE: Image Synthesis by Image-Guided Model Inversion
Pei Wang, Yijun Li, Krishna Kumar Singh, Jingwan Lu, Nuno Vasconcelos; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3681-3690
Synthesizing variations of a specific reference image with semantically valid co ntent is an important task in terms of personalized generation as well as for da ta augmentation. In this work, we propose an inversion based method, denoted as IMAge-Guided model INvErsion (IMAGINE), to generate high-quality and diverse ima ges only from one single training sample. We mainly leverage the knowledge of im age semantics from a pre-trained classifier and achieve plausible generations vi a matching multi-level feature representations in the classifier, associated wit h adversarial training with an external discriminator. IMAGINE enables the synth esis procedure to be able to simultaneously 1) enforce semantic specificity cons traints during the synthesis, 2) produce realistic images without the introducti on of generator training, 3) allow fine controls over the synthesized image, and 4) be model-compact. With extensive experimental results, we demonstrate qualit atively and quantitatively that IMAGINE performs favorably against state-of-the-art GAN-based and inversion-based methods, across three different image domains, i.e., the object, scene and texture.
****
Neural Scene Graphs for Dynamic Scenes
Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, Felix Heide; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 1, pp. 2856-2865
Recent implicit neural rendering methods have demonstrated that it is possible t o learn accurate view synthesis for complex scenes by predicting their volumetri c density and color supervised solely by a set of RGB images. However, existing methods are restricted to learning efficient representations of static scenes th at encode all scene objects into a single neural network, and they lack the abil ity to represent dynamic scenes and decompose scenes into individual objects. In this work, we present the first neural rendering method that represents multi-o bject dynamic scenes as scene graphs. We propose a learned scene graph represent ation, which encodes object transformations and radiance, allowing us to efficie

ntly render novel arrangements and views of the scene. To this end, we learn implicitly encoded scenes, combined with a jointly learned latent representation to describe similar objects with a single implicit function. We assess the proposed method on synthetic and real automotive data, validating that our approach learns dynamic scenes -- only by observing a video of this scene -- and allows for rendering novel photo-realistic views of novel scene compositions with unseen sets of objects at unseen poses.
****
RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words
Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15465-15474
Recent progress on visual question answering has explored the merits of grid features for vision language tasks. Meanwhile, transformer-based models have shown remarkable performance in various sequence prediction problems. However, the spatial information loss of grid features caused by flattening operation, as well as the defect of the transformer model in distinguishing visual words and non visual words, are still left unexplored. In this paper, we first propose Grid-Augmented (GA) module, in which relative geometry features between grids are incorporated to enhance visual representations. Then, we build a BERTbased language model to extract language context and propose Adaptive-Attention (AA) module on top of a transformer decoder to adaptively measure the contribution of visual and language cues before making decisions for word prediction. To prove the generality of our proposals, we apply the two modules to the vanilla transformer model to build our Relationship-Sensitive Transformer (RSTNet) for image captioning task. The proposed model is tested on the MSCOCO benchmark, where it achieves new state-ofart results on both the Karpathy test split and the online test server. Source code is available at GitHub 1.
****
Time Lens: Event-Based Video Frame Interpolation
Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16155-16164
State-of-the-art frame interpolation methods generate intermediate frames by inferring object motions in the image from consecutive key-frames. In the absence of additional information, first-order approximations, i.e. optical flow, must be used, but this choice restricts the types of motions that can be modeled, leading to errors in highly dynamic scenarios. Event cameras are novel sensors that address this limitation by providing auxiliary visual information in the blind-time between frames. They asynchronously measure per-pixel brightness changes and do this with high temporal resolution and low latency. Event-based frame interpolation methods typically adopt a synthesis-based approach, where predicted frame residuals are directly applied to the key-frames. However, while these approaches can capture non-linear motions they suffer from ghosting and perform poorly in low-texture regions with few events. Thus, synthesis-based and flow-based approaches are complementary. In this work, we introduce Time Lens, a novel method that leverages the advantages of both. We extensively evaluate our method on three synthetic and two real benchmarks where we show an up to 5.21 dB improvement in terms of PSNR over state-of-the-art frame-based and event-based methods. Finally, we release a new large-scale dataset in highly dynamic scenarios, aimed at pushing the limits of existing methods.
****
FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space
Quande Liu, Cheng Chen, Jing Qin, Qi Dou, Pheng-Ann Heng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1013-1023
Federated learning allows distributed medical institutions to collaboratively learn a shared prediction model with privacy protection. While at clinical deployment, the models trained in federated learning can still suffer from performance

drop when applied to completely unseen hospitals outside the federation. In this
 paper, we point out and solve a novel problem setting of federated domain gener
alization, which aims to learn a federated model from multiple distributed sourc
e domains such that it can directly generalize to unseen target domains. We pres
ent a novel approach, named as Episodic Learning in Continuous Frequency Space (
ELCFS), for this problem by enabling each client to exploit multi-source data di
stributions under the challenging constraint of data decentralization. Our appro
ach transmits the distribution information across clients in a privacy-protectin
g way through an effective continuous frequency space interpolation mechanism. W
ith the transferred multi-source distributions, we further carefully design a bo
undary-oriented episodic learning paradigm to expose the local learning to domai
n distribution shifts and particularly meet the challenges of model generalizati
on in medical image segmentation scenario. The effectiveness of our method is de
monstrated with superior performance over state-of-the-arts and in-depth ablatio
n experiments on two medical image segmentation tasks. The code is available at
"https://github.com/liuquande/FedDG-ELCFS".
****
Anomaly Detection in Video via Self-Supervised and Multi-Task Learning
Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz K
han, Marius Popescu, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2021, pp. 12742-12752
Anomaly detection in video is a challenging computer vision problem. Due to the
lack of anomalous events at training time, anomaly detection requires the design
 of learning methods without full supervision. In this paper, we approach anomal
ous event detection in video through self-supervised and multi-task learning at
the object level. We first utilize a pre-trained detector to detect objects. The
n, we train a 3D convolutional neural network to produce discriminative anomaly-
specific information by jointly learning multiple proxy tasks: three self-superv
ised and one based on knowledge distillation. The self-supervised tasks are: (i)
 discrimination of forward/backward moving objects (arrow of time), (ii) discrim
ination of objects in consecutive/intermittent frames (motion irregularity) and
(iii) reconstruction of object-specific appearance information. The knowledge di
stillation task takes into account both classification and detection information
, generating large prediction discrepancies between teacher and student models w
hen anomalies occur. To the best of our knowledge, we are the first to approach
anomalous event detection in video as a multi-task learning problem, integrating
 multiple self-supervised and knowledge distillation proxy tasks in a single arc
hitecture. Our lightweight architecture outperforms the state-of-the-art methods
 on three benchmarks: Avenue, ShanghaiTech and UCSD Ped2. Additionally, we perfo
rm an ablation study demonstrating the importance of integrating self-supervised
 learning and normality-specific distillation in a multi-task learning setting.
****
Multiresolution Knowledge Distillation for Anomaly Detection
Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, H
amid R. Rabiee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2021, pp. 14902-14912
Unsupervised representation learning has proved to be a critical component of an
omaly detection/localization in images. The challenges to learn such a represent
ation are two-fold. Firstly, the sample size is not often large enough to learn
a rich generalizable representation through conventional techniques. Secondly, w
hile only normal samples are available at training, the learned features should
be discriminative of normal and anomalous samples. Here, we propose to use the "
distillation" of features at various layers of an expert network, which is pre-t
rained on ImageNet, into a simpler cloner network to tackle both issues. We dete
ct and localize anomalies using the discrepancy between the expert and cloner ne
tworks' intermediate activation values given an input sample. We show that consi
dering multiple intermediate hints in distillation leads to better exploitation
of the expert's knowledge and a more distinctive discrepancy between the two net
works, compared to utilizing only the last layer activation values. Notably, pre
vious methods either fail in precise anomaly localization or need expensive regi

on-based training. In contrast, with no need for any special or intensive traini
ng procedure, we incorporate interpretability algorithms in our novel framework
to localize anomalous regions. Despite the striking difference between some test
 datasets and ImageNet, we achieve competitive or significantly superior results
 compared to SOTA on MNIST, F-MNIST, CIFAR-10, MVTecAD, Retinal-OCT, and two oth
er medical datasets on both anomaly detection and localization.
****
Joint Learning of 3D Shape Retrieval and Deformation
Mikaela Angelina Uy, Vladimir G. Kim, Minhyuk Sung, Noam Aigerman, Siddhartha Ch
audhuri, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2021, pp. 11713-11722
We propose a novel technique for producing high-quality 3D models that match a g
iven target object image or scan. Our method is based on retrieving an existing
shape from a database of 3D models and then deforming its parts to match the tar
get shape. Unlike previous approaches that independently focus on either shape r
etrieval or deformation, we propose a joint learning procedure that simultaneous
ly trains the neural deformation module along with the embedding space used by t
he retrieval module. This enables our network to learn a deformation-aware embed
ding space, so that retrieved models are more amenable to match the target after
 an appropriate deformation. In fact, we use the embedding space to guide the sh
ape pairs used to train the deformation module, so that it invests its capacity
in learning deformations between meaningful shape pairs. Furthermore, our novel
part-aware deformation module can work with inconsistent and diverse part-struct
ures on the source shapes. We demonstrate the benefits of our joint training not
 only on our novel framework, but also on other state-of-the-art neural deformat
ion modules proposed in recent years. Lastly, we also show that our jointly-trai
ned method outperforms various non-joint baselines.
****
Learning Spatially-Variant MAP Models for Non-Blind Image Deblurring
Jiangxin Dong, Stefan Roth, Bernt Schiele; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4886-4895
The classical maximum a-posteriori (MAP) framework for non-blind image deblurrin
g requires defining suitable data and regularization terms, whose interplay yiel
ds the desired clear image through optimization. The vast majority of prior work
 focuses on advancing one of these two crucial ingredients, while keeping the ot
her one standard. Considering the indispensable roles and interplay of both data
 and regularization terms, we propose a simple and effective approach to jointly
 learn these two terms, embedding deep neural networks within the constraints of
 the MAP framework, trained in an end-to-end manner. The neural networks not onl
y yield suitable image-adaptive features for both terms, but actually predict pe
r-pixel spatially-variant features instead of the commonly used spatially-unifor
m ones. The resulting spatially-variant data and regularization terms particular
ly improve the restoration of fine-scale structures and detail. Quantitative and
 qualitative results underline the effectiveness of our approach, substantially
outperforming the current state of the art.
****
FCPose: Fully Convolutional Multi-Person Pose Estimation With Dynamic Instance-A
ware Convolutions
Weian Mao, Zhi Tian, Xinlong Wang, Chunhua Shen; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9034-9043
We propose a fully convolutional multi-person pose estimation framework using dy
namic instance-aware convolutions, termed FCPose. Different from existing method
s, which often require ROI (Region of Interest) operations and/or grouping post-
processing, FCPose eliminates the ROIs and grouping post-processing with dynamic
 instance-aware keypoint estimation heads. The dynamic keypoint heads are condit
ioned on each instance (person), and can encode the instance concept in the dyna
mically-generated weights of their filters. Moreover, with the strong representa
tion capacity of dynamic convolutions, the keypoint heads in FCPose are designed
 to be very compact, resulting in fast inference and makes FCPose have almost co
nstant inference time regardless of the number of persons in the image. For exam

ple, on the COCO dataset, a real-time version of FCPose using the DLA-34 backbon
e infers about 4.5 times faster than Mask R-CNN (ResNet-101) (41.67 FPS vs. 9.26
 FPS) while achieving improved performance (64.8% AP vs. 64.3% AP). FCPose also
offers better speed/accuracy trade-off than other state-of-the-art methods. Our
experiment results show that FCPose is a simple yet effective multi-person pose
estimation framework. Code is available at: https://git.io/AdelaiDet
****
BoxInst: High-Performance Instance Segmentation With Box Annotations
Zhi Tian, Chunhua Shen, Xinlong Wang, Hao Chen; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5443-5452
We present a high-performance method that can achieve mask-level instance segmen
tation with only bounding-box annotations for training. While this setting has b
een studied in the literature, here we show significantly stronger performance w
ith a simple design (e.g., dramatically improving previous best reported mask AP
 of 21.1% to 31.6% on the COCO dataset). Our core idea is to redesign the loss o
f learning masks in instance segmentation, with no modification to the segmentat
ion network itself. The new loss functions can supervise the mask training witho
ut relying on mask annotations. This is made possible with two loss terms, namel
y, 1) a surrogate term that minimizes the discrepancy between the projections of
 the ground-truth box and the predicted mask; 2) a pairwise loss that can exploi
t the prior that proximal pixels with similar colors are very likely to have the
 same category label. Experiments demonstrate that the redesigned mask loss can
yield surprisingly high-quality instance masks with only box annotations. For ex
ample, without using any mask annotations, with a ResNet-101 backbone and 3x tra
ining schedule, we achieve 33.2% mask AP on COCO test-dev split (vs. 39.1% of th
e fully supervised counterpart). Our excellent experiment results on COCO and Pa
scal VOC indicate that our method dramatically narrows the performance gap betwe
en weakly and fully supervised instance segmentation. Code is available at https
://git.io/AdelaiDet
****
Modeling Multi-Label Action Dependencies for Temporal Action Localization
Praveen Tirupattur, Kevin Duarte, Yogesh S Rawat, Mubarak Shah; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
pp. 1460-1470
Real world videos contain many complex actions with inherent relationships betwe
en action classes. In this work, we propose an attention-based architecture that
 model these action relationships for the task of temporal action localization i
n untrimmed videos. As opposed to previous works which leverage video-level co-o
ccurrence of actions, we distinguish the relationships between actions that occu
r at the same time-step and actions that occur at different time-steps (i.e. tho
se which precede or follow each other). We define these distinct relationships a
s action dependencies. We propose to improve action localization performance by
modeling these action dependencies in a novel attention based Multi-Label Action
 Dependency (MLAD) layer. The MLAD layer consists of two branches: a Co-occurren
ce Dependency Branch and a Temporal Dependency Branch to model co-occurrence act
ion dependencies and temporal action dependencies, respectively. We observe that
 existing metrics used for multi-label classification do not explicitly measure
how well action dependencies are modeled, therefore, we propose novel metrics wh
ich consider both co-occurrence and temporal dependencies between action classes
. Through empirical evaluation and extensive analysis we show improved performan
ce over state-of-the art methods on multi-label action localization benchmarks (
MultiTHUMOS and Charades) in terms of f-mAP and our proposed metric.
****
HCRF-Flow: Scene Flow From Point Clouds With Continuous High-Order CRFs and Posi
tion-Aware Flow Embedding
Ruibo Li, Guosheng Lin, Tong He, Fayao Liu, Chunhua Shen; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 36
4-373
Scene flow in 3D point clouds plays an important role in understanding dynamic e
nvironments. Although significant advances have been made by deep neural network

s, the performance is far from satisfactory as only per-point translational moti
on is considered, neglecting the constraints of the rigid motion in local region
s. To address the issue, we propose to introduce the motion consistency to force
 the smoothness among neighboring points. In addition, constraints on the rigidi
ty of the local transformation are also added by sharing unique rigid motion par
ameters for all points within each local region. To this end, a high-order CRFs
based relation module (Con-HCRFs) is deployed to explore both point-wise smoothn
ess and region-wise rigidity. To empower the CRFs to have a discriminative unary
 term, we also introduce a position-aware flow estimation module to be incorpora
ted into the Con-HCRFs. Comprehensive experiments on FlyingThings3D and KITTI sh
ow that our proposed framework (HCRF-Flow) achieves state-of-the-art performance
 and significantly outperforms previous approaches substantially.
****
Lite-HRNet: A Lightweight High-Resolution Network
Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, Jingdong Wa
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2021, pp. 10440-10450
We present an efficient high-resolution network, Lite-HRNet, for human pose esti
mation. We start by simply applying the efficient shuffle block in ShuffleNet to
 HRNet (high-resolution network), yielding stronger performance over popular lig
htweight networks, such as MobileNet, ShuffleNet, and Small HRNet. We find that
the heavily-used pointwise (1x1) convolutions in shuffle blocks become the compu
tational bottleneck. We introduce a lightweight unit, conditional channel weight
ing, to replace costly pointwise (1x1) convolutions in shuffle blocks. The compl
exity of channel weighting is linear w.r.t the number of channels and lower than
 the quadratic time complexity for pointwise convolutions. Our solution learns t
he weights from all the channels and over multiple resolutions that are readily
available in the parallel branches in HRNet. It uses the weights as the bridge t
o exchange information across channels and resolutions, compensating the role pl
ayed by the pointwise (1x1) convolution. Lite-HRNet demonstrates superior result
s on human pose estimation over popular lightweight networks. Moreover, Lite-HRN
et can be easily applied to semantic segmentation task in the same lightweight m
anner. The code and models have been publicly available at https://github.com/HR
Net/Lite-HRNet.
****
Self-Supervised Video Representation Learning by Context and Motion Decoupling
Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, Rong Jin; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021,
 pp. 13886-13895
A key challenge in self-supervised video representation learning is how to effec
tively capture motion information besides context bias. While most existing work
s implicitly achieve this with video-specific pretext tasks (e.g., predicting cl
ip orders, time arrows, and paces), we develop a method that explicitly decouple
s motion supervision from context bias through a carefully designed pretext task
. Specifically, we take the key frames and motion vectors in compressed videos (
e.g., in H.264 format) as the supervision sources for context and motion, respec
tively, which can be efficiently extracted at over 500 fps on CPU. Then we desig
n two pretext tasks that are jointly optimized: a context matching task where a
pairwise contrastive loss is cast between video clip and key frame features; and
 a motion prediction task where clip features, passed through an encoder-decoder
 network, are used to estimate motion features in a near future. These two tasks
 use a shared video backbone and separate MLP heads. Experiments show that our a
pproach improves the quality of the learned video representation over previous w
orks, where we obtain absolute gains of 16.0% and 11.1% in video retrieval recal
l on UCF101 and HMDB51, respectively. Moreover, we find the motion prediction to
 be a strong regularization for video networks, where using it as an auxiliary t
ask improves the accuracy of action recognition with a margin of 7.4%   13.8%.
****
ReAgent: Point Cloud Registration Using Imitation and Reinforcement Learning
Dominik Bauer, Timothy Patten, Markus Vincze; Proceedings of the IEEE/CVF Confer

ence on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14586-14594
Point cloud registration is a common step in many 3D computer vision tasks such as object pose estimation, where a 3D model is aligned to an observation. Classical registration methods generalize well to novel domains but fail when given a noisy observation or a bad initialization. Learning-based methods, in contrast, are more robust but lack in generalization capacity. We propose to consider iterative point cloud registration as a reinforcement learning task and, to this end, present a novel registration agent (ReAgent). We employ imitation learning to initialize its discrete registration policy based on a steady expert policy. Integration with policy optimization, based on our proposed alignment reward, further improves the agent's registration performance. We compare our approach to classical and learning-based registration methods on both ModelNet40 (synthetic) and ScanObjectNN (real data) and show that our ReAgent achieves state-of-the-art accuracy. The lightweight architecture of the agent, moreover, enables reduced inference time as compared to related approaches.
****
Uncertainty Guided Collaborative Training for Weakly Supervised Temporal Action Detection
Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 53-63
Weakly supervised temporal action detection aims to localize temporal boundaries of actions and identify their categories simultaneously with only video-level category labels during training. Among existing methods, attention-based methods have achieved superior performance by separating action and non-action segments. However, without the segment-level ground-truth supervision, the quality of the attention weight hinders the performance of these methods. To alleviate this problem, we propose a novel Uncertainty Guided Collaborative Training (UGCT) strategy, which mainly includes two key designs: (1) The first design is an online pseudo label generation module, in which the RGB and FLOW streams work collaboratively to learn from each other. (2) The second design is an uncertainty aware learning module, which can mitigate the noise in the generated pseudo labels. These two designs work together to promote the model performance effectively and efficiently. Experimental results on three state-of-the-art attentionbased methods demonstrate that the proposed training strategy can significantly improve the performance of these methods, e.g., more than 4% for all three methods in terms of mAP@IoU=0.5 on the THUMOS14 dataset.
****
Dynamic Probabilistic Graph Convolution for Facial Action Unit Intensity Estimation
Tengfei Song, Zijun Cui, Yuru Wang, Wenming Zheng, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4845-4854
Deep learning methods have been widely applied to automatic facial action unit (AU) intensity estimation and achieved state-of-the-art performance. These methods, however, are mostly appearance-based and fail to exploit the underlying structural information among the AUs. In this paper, we propose a novel dynamic probabilistic graph convolution (DPG) model to simultaneously exploit AU appearances, AU dynamics, and their semantic structural dependencies for AU intensity estimation. First, we propose to use Bayesian Network to capture the inherent dependencies among the AUs. Second, we introduce probabilistic graph convolution that allows to perform graph convolution on the distribution of Bayesian Network structure to extract AU structural features. Finally, we introduce a dynamic deep model based on LSTM to simultaneously combine AU appearance features, AU dynamic features, and AU structural features for improved AU intensity estimation. In experiments, our method achieves comparable and even better performance with state-of-the-art methods on two benchmark facial AU intensity estimation databases, i.e., FERA 2015 and DISFA.
****
Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is Al

l You Need?
Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ay
ed, Jose Dolz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2021, pp. 13979-13988
We show that the way inference is performed in few-shot segmentation tasks has a
 substantial effect on performances--an aspect often overlooked in the literatur
e in favor of the meta-learning paradigm. We introduce a transductive inference
for a given query image, leveraging the statistics of its unlabeled pixels, by o
ptimizing a new loss containing three complementary terms: i) the cross-entropy
on the labeled support pixels; ii) the Shannon entropy of the posteriors on the
unlabeled query image pixels; and iii) a global KL-divergence regularizer based
on the proportion of the predicted foreground. As our inference uses a simple li
near classifier of the extracted features, its computational load is comparable
to inductive inference and can be used on top of any base training. Foregoing ep
isodic training and using only standard cross-entropy training on the base class
es, our inference yields competitive performances on standard benchmarks in the
1-shot scenarios. As the number of available shots increases, the gap in perform
ances widens: on PASCAL-5i, our method brings about 5% and 6% improvements over
the state-of-the-art, in the 5- and 10-shot scenarios, respectively. Furthermore
, we introduce a new setting that includes domain shifts, where the base and nov
el classes are drawn from different datasets. Our method achieves the best perfo
rmances in this more realistic setting. Our code is freely available online: htt
ps://github.com/mboudiaf/RePRI-for-Few-Shot-Segmentation.
****
Spatial-Temporal Correlation and Topology Learning for Person Re-Identification
in Videos
Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, Qibin Sun; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp.
 4370-4379
Video-based person re-identification aims to match pedestrians from video sequen
ces across non-overlapping camera views. The key factor for video person re-iden
tification is to effectively exploit both spatial and temporal clues from video
sequences. In this work, we propose a novel Spatial-Temporal Correlation and Top
ology Learning framework (CTL) to pursue discriminative and robust representatio
n by modeling cross-scale spatial-temporal correlation. Specifically, CTL utiliz
es a CNN backbone and a key-points estimator to extract semantic local features
from human body at multiple granularities as graph nodes. It explores a context-
reinforced topology to construct multi-scale graphs by considering both global c
ontextual information and physical connections of human body. Moreover, a 3D gra
ph convolution and a cross-scale graph convolution are designed, which facilitat
e direct cross-spacetime and cross-scale information propagation for capturing h
ierarchical spatial-temporal dependencies and structural information. By jointly
 performing the two convolutions, CTL effectively mines comprehensive clues that
 are complementary with appearance information to enhance representational capac
ity. Extensive experiments on two video benchmarks have demonstrated the effecti
veness of the proposed method and the state-of-the-art performance.
****
SPSG: Self-Supervised Photometric Scene Generation From RGB-D Scans
Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, Matthias Niessner; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2021, pp. 1747-1756
We present SPSG, a novel approach to generate high-quality, colored 3D models of
 scenes from RGB-D scan observations by learning to infer unobserved scene geome
try and color in a self-supervised fashion. Our self-supervised approach learns
to jointly inpaint geometry and color by correlating an incomplete RGB-D scan wi
th a more complete version of that scan. Notably, rather than relying on 3D reco
nstruction losses to inform our 3D geometry and color reconstruction, we propose
 adversarial and perceptual losses operating on 2D renderings in order to achiev
e high-resolution, high-quality colored reconstructions of scenes. This exploits
 the high-resolution, self-consistent signal from individual raw RGB-D frames, i

n contrast to fused 3D reconstructions of the frames which exhibit inconsistenci es from view-dependent effects, such as color balancing or pose inconsistencies. Thus, by informing our 3D scene generation directly through 2D signal, we produ ce high-quality colored reconstructions of 3D scenes, outperforming state of the art on both synthetic and real data.
****
Neural Auto-Exposure for High-Dynamic Range Object Detection
Emmanuel Onzon, Fahim Mannan, Felix Heide; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7710-7720
Real-world scenes have a dynamic range of up to 280 dB that today's imaging sens ors cannot directly capture. Existing live vision pipelines tackle this fundamen tal challenge by relying on high dynamic range (HDR) sensors that try to recover HDR images from multiple captures with different exposures. While HDR sensors s ubstantially increase the dynamic range, they are not without disadvantages, inc luding severe artifacts for dynamic scenes, reduced fill-factor, lower resolutio n, and high sensor cost. At the same time, traditional auto-exposure methods for low-dynamic range sensors have advanced as proprietary methods relying on image statistics separated from downstream vision algorithms. In this work, we revisi t auto-exposure control as an alternative to HDR sensors. We propose a neural ne twork for exposure selection that is trained jointly, end-to-end with an object detector and an image signal processing (ISP) pipeline. To this end, we use an H DR dataset for automotive object detection and an HDR training procedure. We val idate that the proposed neural auto-exposure control, which is tailored to objec t detection, outperforms conventional auto-exposure methods by more than 6 point s in mean average precision (mAP).
****
Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Tr ansformers
Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, Li Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6881-6890
Most recent semantic segmentation methods adopt a fully-convolutional network (F CN) with an encoder-decoder architecture. The encoder progressively reduces the spatial resolution and learns more abstract/semantic visual concepts with larger receptive fields. Since context modeling is critical for segmentation, the late st efforts have been focused on increasing the receptive field, through either d ilated/atrous convolutions or inserting attention modules. However, the encoder-decoder based FCN architecture remains unchanged. In this paper, we aim to provi de an alternative perspective by treating semantic segmentation as a sequence-to -sequence prediction task. Specifically, we deploy a pure transformer (i.e., wit hout convolution and resolution reduction) to encode an image as a sequence of p atches. With the global context modeled in every layer of the transformer, this encoder can be combined with a simple decoder to provide a powerful segmentation model, termed SEgmentation TRansformer (SETR). Extensive experiments show that SETR achieves new state of the art on ADE20K (50.28% mIoU), Pascal Context (55.8 3% mIoU) and competitive results on Cityscapes. Particularly, we achieve the fir st position in the highly competitive ADE20K test server leaderboard on the day of submission.
****
Interpreting Super-Resolution Networks With Local Attribution Maps
Jinjin Gu, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9199-9208
Image super-resolution (SR) techniques have been developing rapidly, benefiting from the invention of deep networks and its successive breakthroughs. However, i t is acknowledged that deep learning and deep neural networks are difficult to i nterpret. SR networks inherit this mysterious nature and little works make attem pt to understand them. In this paper, we perform attribution analysis of SR netw orks, which aims at finding the input pixels that strongly influence the SR resu lts. We propose a novel attribution approach called local attribution map (LAM),

which inherits the integral gradient method yet with two unique features. One is to use the blurred image as the baseline input, and the other is to adopt the progressive blurring function as the path function. Based on LAM, we show that: (1) SR networks with a wider range of involved input pixels could achieve better performance. (2) Attention networks and non-local networks extract features from a wider range of input pixels. (3) Comparing with the range that actually contributes, the receptive field is large enough for most deep networks. (4) For SR networks, textures with regular stripes or grids are more likely to be noticed, while complex semantics are difficult to utilize. Our work opens new directions for designing SR networks and interpreting low-level vision deep models.
****
Multi-Target Domain Adaptation With Collaborative Consistency Learning
Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8187-8196
Recently unsupervised domain adaptation for the semantic segmentation task has become more and more popular due to the high-cost of pixel-level annotation on real-world images. However, most domain adaptation methods are only restricted to single-source-single-target pair, and can not be directly extended to multiple target domains. In this work, we propose a collaborative learning framework to achieve unsupervised multi-target domain adaptation. An unsupervised domain adaptation expert model is first trained for each source-target pair and is further encouraged to collaborate with each other through a bridge built between different target domains. These expert models are further improved by adding the regularization of making the consistent pixel-wise prediction for each sample with the same structured context. To obtain a single model that works across multiple target domains, we propose to simultaneously learn a student model which is trained to not only imitate the output of each expert on the corresponding target domain but also to pull different expert close to each other with regularization on their weights. Extensive experiments demonstrate that the proposed method can effectively exploit rich structured information contained in both labeled source domain and multiple unlabeled target domains. Not only does it perform well across multiple target domains but also performs favorably against state-of-the-art unsupervised domain adaptation methods specially trained on a single source-target pair.
****
Troubleshooting Blind Image Quality Models in the Wild
Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, Kede Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16256-16265
Recently, the group maximum differentiation competition (gMAD) has been used to improve blind image quality assessment (BIQA) models, with the help of full-reference metrics. When applying this type of approach to troubleshoot "best-performing" BIQA models in the wild, we are faced with a practical challenge: it is highly nontrivial to obtain stronger competing models for efficient failure-spotting. Inspired by recent findings that difficult samples of deep models may be exposed through network pruning, we construct a set of "self-competitors," as random ensembles of pruned versions of the target model to be improved. Diverse failures can then be efficiently identified via self-gMAD competition. Next, we fine-tune both the target and its pruned variants on the human-rated gMAD set. This allows all models to learn from their respective failures, preparing themselves for the next round of self-gMAD competition. Experimental results demonstrate that our method efficiently troubleshoots BIQA models in the wild with improved generalizability.
****
Semantic Palette: Guiding Scene Generation With Class Proportions
Guillaume Le Moing, Tuan-Hung Vu, Himalaya Jain, Patrick Perez, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9342-9350
Despite the recent progress of generative adversarial networks (GANs) at synthes

izing photo-realistic images, producing complex urban scenes remains a challenging problem. Previous works break down scene generation into two consecutive phases: unconditional semantic layout synthesis and image synthesis conditioned on layouts. In this work, we propose to condition layout generation as well for higher semantic control: given a vector of class proportions, we generate layouts with matching composition. To this end, we introduce a conditional framework with novel architecture designs and learning objectives, which effectively accommodates class proportions to guide the scene generation process. The proposed architecture also allows partial layout editing with interesting applications. Thanks to the semantic control, we can produce layouts close to the real distribution, helping enhance the whole scene generation process. On different metrics and urban scene benchmarks, our models outperform existing baselines. Moreover, we demonstrate the merit of our approach for data augmentation: semantic segmenters trained on real layout-image pairs along with additional ones generated by our approach outperform models only trained on real pairs.
****
Physics-Based Iterative Projection Complex Neural Network for Phase Retrieval in Lensless Microscopy Imaging
Feilong Zhang, Xianming Liu, Cheng Guo, Shiyi Lin, Junjun Jiang, Xiangyang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10523-10531
Phase retrieval from intensity-only measurements plays a central role in many real-world imaging tasks. In recent years, deep neural networks based methods emerge and show promising performance for phase retrieval. However, their interpretability and generalization still remain a major challenge. In this paper, we propose to combine the advantages of both model-based alternative projection method and deep neural network for phase retrieval, so as to achieve network interpretability and inference effectiveness simultaneously. Specifically, we unfold the iterative process of the alternative projection phase retrieval into a feed-forward neural network, whose layers mimic the processing flow. The physical model of the imaging process is then naturally embedded into the neural network structure. Moreover, a complex-valued U-Net is proposed for defining image priori for forward and backward projection in dual planes. Finally, we designate physics-based formulation as an untrained deep neural network, whose weights are enforced to fit to the given intensity measurements. In summary, our scheme for phase retrieval is effective, interpretable, physics-based and unsupervised. Experimental results demonstrate that our method achieves superior performance compared with the state-of-the-arts in a practical phase retrieval application---lensless microscopy imaging.
****
Causal Attention for Vision-Language Tasks
Xu Yang, Hanwang Zhang, Guojun Qi, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9847-9857
We present a novel attention mechanism: Causal Attention (CATT), to remove the ever-elusive confounding effect in existing attention-based vision-language models. This effect causes harmful bias that misleads the attention module to focus on the spurious correlations in training data, damaging the model generalization. As the confounder is unobserved in general, we use the front-door adjustment to realize the causal intervention, which does not require any knowledge on the confounder. Specifically, CATT is implemented as a combination of 1) In-Sample Attention (IS-ATT) and 2) Cross-Sample Attention (CS-ATT), where the latter forcibly brings other samples into every IS-ATT, mimicking the causal intervention. CATT abides by the Q-K-V convention and hence can replace any attention module such as top-down attention and self-attention in Transformers. CATT improves various popular attention-based vision-language models by considerable margins. In particular, we show that CATT has great potential in large-scale pre-training, e.g., it can promote the lighter LXMERT [??], which uses fewer data and less computational power, comparable to the heavier UNITER [??]. Code is published in https://github.com/yangxuntu/lxmertcatt.
****

Scene Text Telescope: Text-Focused Scene Image Super-Resolution
Jingye Chen, Bin Li, Xiangyang Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12026-12035
Image super-resolution, which is often regarded as a preprocessing procedure of scene text recognition, aims to recover the realistic features from a low-resolution text image. It has always been challenging due to large variations in text shapes, fonts, backgrounds, etc. However, most existing methods employ generic super-resolution frameworks to handle scene text images while ignoring text-specific properties such as text-level layouts and character-level details. In this paper, we establish a text-focused super-resolution framework, called Scene Text Telescope (STT). In terms of text-level layouts, we propose a Transformer-Based Super-Resolution Network (TBSRN) containing a Self-Attention Module to extract sequential information, which is robust to tackle the texts in arbitrary orientations. In terms of character-level details, we propose a Position-Aware Module and a Content-Aware Module to highlight the position and the content of each character. By observing that some characters look indistinguishable in low-resolution conditions, we use a weighted cross-entropy loss to tackle this problem. We conduct extensive experiments, including text recognition with pre-trained recognizers and image quality evaluation, on TextZoom and several scene text recognition benchmarks to assess the super-resolution images. The experimental results show that our STT can indeed generate text-focused super-resolution images and outperform the existing methods in terms of recognition accuracy.
****
NeuTex: Neural Texture Mapping for Volumetric Neural Rendering
Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7119-7128
Recent work has demonstrated that volumetric scene representations combined with differentiable volume rendering can enable photo-realistic rendering for challenging scenes that mesh reconstruction fails on. However, these methods entangle geometry and appearance in a ""black-box"" volume that cannot be edited. Instead, we present an approach that explicitly disentangles geometry--represented as a continuous 3D volume--from appearance--represented as a continuous 2D texture map. We achieve this by introducing a 3D-to-2D texture mapping (or surface parameterization) network into volumetric representations. We constrain this texture mapping network using an additional 2D-to-3D inverse mapping network and a novel cycle consistency loss to make 3D surface points map to 2D texture points that map back to the original 3D points. We demonstrate that this representation can be reconstructed using only multi-view image supervision and generates high-quality rendering results. More importantly,by separating geometry and texture, we allow users to edit appearance by simply editing 2D texture maps.
****
Improving Calibration for Long-Tailed Recognition
Zhisheng Zhong, Jiequan Cui, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16489-16498
Deep neural networks may perform poorly when training datasets are heavily class-imbalanced. Recently, two-stage methods decouple representation learning and classifier learning to improve performance. But there is still the vital issue of miscalibration. To address it, we design two methods to improve calibration and performance in such scenarios. Motivated by the fact that predicted probability distributions of classes are highly related to the numbers of class instances, we propose label-aware smoothing to deal with different degrees of over-confidence for classes and improve classifier learning. For dataset bias between these two stages due to different samplers, we further propose shifted batch normalization in the decoupling framework. Our proposed methods set new records on multiple popular long-tailed recognition benchmark datasets, including CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, Places-LT, and iNaturalist 2018.
****
Learning Affinity-Aware Upsampling for Deep Image Matting
Yutong Dai, Hao Lu, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Comp

uter Vision and Pattern Recognition (CVPR), 2021, pp. 6841-6850
We show that learning affinity in upsampling provides an effective and efficient approach to exploit pairwise interactions in deep networks. Second-order features are commonly used in dense prediction to build adjacent relations with a learnable module after upsampling such as non-local blocks. Since upsampling is essential, learning affinity in upsampling can avoid additional propagation layers, offering the potential for building compact models. By looking at existing upsampling operators from a unified mathematical perspective, we generalize them into a second-order form and introduce Affinity-Aware Upsampling (A2U) where upsampling kernels are generated using a light-weight low-rank bilinear model and are conditioned on second-order features. Our upsampling operator can also be extended to downsampling. We discuss alternative implementations of A2U and verify their effectiveness on two detail-sensitive tasks: image reconstruction on a toy dataset; and a large-scale image matting task where affinity-based ideas constitute mainstream matting approaches. In particular, results on the Composition-1k matting dataset show that A2U achieves a 14% relative improvement in the SAD metric against a strong baseline with negligible increase of parameters (< 0.5%). Compared with the state-of-the-art matting network, we achieve 8% higher performance with only 40% model complexity.
****
Improving Multiple Pedestrian Tracking by Track Management and Occlusion Handling
Daniel Stadler, Jurgen Beyerer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10958-10967
Multi-pedestrian trackers perform well when targets are clearly visible making the association task quite easy. However, when heavy occlusions are present, a mechanism to reidentify persons is needed. The common approach is to extract visual features from new detections and compare them with the features of previously found tracks. Since those detections can have substantial overlaps with nearby targets - especially in crowded scenarios - the extracted features are insufficient for a reliable re-identification. In contrast, we propose a novel occlusion handling strategy that explicitly models the relation between occluding and occluded tracks outperforming the feature-based approach, while not depending on a separate re-identification network. Furthermore, we improve the track management of a regression-based method in order to bypass missing detections and to deal with tracks leaving the scene at the border of the image. Finally, we apply our tracker in both temporal directions and merge tracklets belonging to the same target, which further enhances the performance. We demonstrate the effectiveness of our tracking components with ablative experiments and surpass the state-of-the-art methods on the three popular pedestrian tracking benchmarks MOT16, MOT17, and MOT20.
****
Revamping Cross-Modal Recipe Retrieval With Hierarchical Transformers and Self-Supervised Learning
Amaia Salvador, Erhan Gundogdu, Loris Bazzani, Michael Donoser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15475-15484
Cross-modal recipe retrieval has recently gained substantial attention due to the importance of food in people's lives, as well as the availability of vast amounts of digital cooking recipes and food images to train machine learning models. In this work, we revisit existing approaches for cross-modal recipe retrieval and propose a simplified end-to-end model based on well established and high performing encoders for text and images. We introduce a hierarchical recipe Transformer which attentively encodes individual recipe components (titles, ingredients and instructions). Further, we propose a self-supervised loss function computed on top of pairs of individual recipe components, which is able to leverage semantic relationships within recipes, and enables training using both image-recipe and recipe-only samples. We conduct a thorough analysis and ablation studies to validate our design choices. As a result, our proposed method achieves state-of-the-art performance in the cross-modal recipe retrieval task on the Recipe1M data

set. We make code and models publicly available.
****

Geo-FARM: Geodesic Factor Regression Model for Misaligned Pre-Shape Responses in Statistical Shape Analysis

Chao Huang, Anuj Srivastava, Rongjie Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11496-11505

The problem of using covariates to predict shapes of objects in a regression setting is important in many fields. A formal statistical approach, termed geodesic regression model, is commonly used for modeling and analyzing relationships between Euclidean predictors and shape responses. Despite its popularity, this model faces several key challenges, including (i) misalignment of shapes due to pre-processing steps, (ii) difficulties in shape alignment due to imaging heterogeneity, and (iii) lack of spatial correlation in shape structures. This paper proposes a comprehensive geodesic factor regression model that addresses all these challenges. Instead of using shapes as extracted from pre-registered data, it takes a more fundamental approach, incorporating alignment step within the proposed regression model and learns them using both pre-shape and covariate data. Additionally, it specifies spatial correlation structures using low-dimensional representations, including latent factors on the tangent space and isotropic error terms. The proposed framework results in substantial improvements in regression performance, as demonstrated through simulation studies and a real data analysis on Corpus Callosum contour data obtained from the ADNI study.
****

MOST: A Multi-Oriented Scene Text Detector With Localization Refinement

Minghang He, Minghui Liao, Zhibo Yang, Humen Zhong, Jun Tang, Wenqing Cheng, Cong Yao, Yongpan Wang, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8813-8822

Over the past few years, the field of scene text detection has progressed rapidly that modern text detectors are able to hunt text in various challenging scenarios. However, they might still fall short when handling text instances of extreme aspect ratios and varying scales. To tackle such difficulties, we propose in this paper a new algorithm for scene text detection, which puts forward a set of strategies to significantly improve the quality of text localization. Specifically, a Text Feature Alignment Module (TFAM) is proposed to dynamically adjust the receptive fields of features based on initial raw detections; a Position-Aware Non-Maximum Suppression (PA-NMS) module is devised to selectively concentrate on reliable raw detections and exclude unreliable ones; besides, we propose an Instance-wise IoU loss for balanced training to deal with text instances of different scales. An extensive ablation study demonstrates the effectiveness and superiority of the proposed strategies. The resulting text detection system, which integrates the proposed strategies with a leading scene text detector EAST, achieves state-of-the-art or competitive performance on various standard benchmarks for text detection while keeping a fast running speed.
****

A Functional Approach to Rotation Equivariant Non-Linearities for Tensor Field Networks.

Adrien Poulenard, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13174-13183

Learning pose invariant representation is a fundamental problem in shape analysis. Most existing deep learning algorithms for 3D shape analysis are not robust to rotations and are often trained on synthetic datasets consisting of pre-aligned shapes, yielding poor generalization to unseen poses. This observation motivates a growing interest in rotation invariant and equivariant methods. The field of rotation equivariant deep learning is developing in recent years thanks to a well established theory of Lie group representations and convolutions. A fundamental problem in equivariant deep learning is to design activation functions which are both informative and preserve equivariance. The recently introduced Tensor Field Network (TFN) framework provides a rotation equivariant network design for point cloud analysis. TFN features undergo a rotation in feature space given a rotation of the input pointcloud. TFN and similar designs consider nonlinearitie

s which operate only over rotation invariant features such as the norm of equiva
riant features to preserve equivariance, making them unable to capture the direc
tional information. In a recent work entitled "Gauge Equivariant Mesh CNNs: Anis
otropic Convolutions on Geometric Graphs" Hann et al. interpret 2D rotation equi
variant features as Fourier coefficients of functions on the circle. In this wor
k we transpose the idea of Hann et al. to 3D by interpreting TFN features as sph
erical harmonics coefficients of functions on the sphere. We introduce a new equ
ivariant nonlinearity and pooling for TFN. We show improvments over the original
 TFN design and other equivariant nonlinearities in classification and segmentat
ion tasks. Furthermore our method is competitive with state of the art rotation
invariant methods in some instances.
****
Leveraging Large-Scale Weakly Labeled Data for Semi-Supervised Mass Detection in
 Mammograms
Yuxing Tang, Zhenjie Cao, Yanbo Zhang, Zhicheng Yang, Zongcheng Ji, Yiwei Wang,
Mei Han, Jie Ma, Jing Xiao, Peng Chang; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3855-3864
Mammographic mass detection is an integral part of a computer-aided diagnosis sy
stem. Annotating a large number of mammograms at pixel-level in order to train a
 mass detection model in a fully supervised fashion is costly and time-consuming
. This paper presents a novel self-training framework for semi-supervised mass d
etection with soft image-level labels generated from diagnosis reports by Mammo-
RoBERTa, a RoBERTa-based natural language processing model fine-tuned on the ful
ly labeled data and associated mammography reports. Starting with a fully superv
ised model trained on the data with pixel-level masks, the proposed framework it
eratively refines the model itself using the entire weakly labeled data (image-l
evel soft label) in a self-training fashion. A novel sample selection strategy i
s proposed to identify those most informative samples for each iteration, based
on the current model output and the soft labels of the weakly labeled data. A so
ft cross-entropy loss and a soft focal loss are also designed to serve as the im
age-level and pixel-level classification loss respectively. Our experiment resul
ts show that the proposed semi-supervised framework can improve the mass detecti
on accuracy on top of the supervised baseline, and outperforms the previous stat
e-of-the-art semi-supervised approaches with weakly labeled data, in some cases
by a large margin.
****
Fast and Accurate Model Scaling
Piotr Dollar, Mannat Singh, Ross Girshick; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 924-932
In this work we analyze strategies for convolutional neural network scaling; tha
t is, the process of scaling a base convolutional network to endow it with great
er computational complexity and consequently representational power. Example sca
ling strategies may include increasing model width, depth, resolution, etc. Whil
e various scaling strategies exist, their tradeoffs are not fully understood. Ex
isting analysis typically focuses on the interplay of accuracy and flops (floati
ng point operations). Yet, as we demonstrate, various scaling strategies affect
model parameters, activations, and consequently actual runtime quite differently
. In our experiments we show the surprising result that numerous scaling strateg
ies yield networks with similar accuracy but with widely varying properties. Thi
s leads us to propose a simple fast compound scaling strategy that encourages pr
imarily scaling model width, while scaling depth and resolution to a lesser exte
nt. Unlike currently popular scaling strategies, which result in about O(s) incr
ease in model activation w.r.t. scaling flops by a factor of s, the proposed fas
t compound scaling results in close to O(sqrt s) increase in activations, while
 achieving excellent accuracy. Fewer activations leads to speedups on modern mem
ory-bandwidth limited hardware (e.g., GPUs). More generally, we hope this work p
rovides a framework for analyzing scaling strategies under various computational
 constraints.
****
Real-Time Sphere Sweeping Stereo From Multiview Fisheye Images

Andreas Meuleman, Hyeonjoong Jang, Daniel S. Jeon, Min H. Kim; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p
p. 11423-11432
A set of cameras with fisheye lenses have been used to capture a wide field of v
iew. The traditional scan-line stereo algorithms based on epipolar geometry are
directly inapplicable to this non-pinhole camera setup due to optical characteri
stics of fisheye lenses; hence, existing complete 360-deg. RGB-D imaging systems
 have rarely achieved real-time performance yet. In this paper, we introduce an
efficient sphere-sweeping stereo that can run directly on multiview fisheye imag
es without requiring additional spherical rectification. Our main contributions
are: First, we introduce an adaptive spherical matching method that accounts for
 each input fisheye camera's resolving power concerning spherical distortion. Se
cond, we propose a fast inter-scale bilateral cost volume filtering method that
refines distance in noisy and textureless regions with the optimal complexity of
 O(n). It enables real-time dense distance estimation while preserving edges. La
stly, the fisheye color and distance images are seamlessly combined into a compl
ete 360-deg. RGB-D image via fast inpainting of the dense distance map. We demon
strate an embedded 360-deg. RGB-D imaging prototype composed of a mobile GPU and
 four fisheye cameras. Our prototype is capable of capturing complete 360-deg. R
GB-D videos with a resolution of two megapixels at 29 fps. Results demonstrate t
hat our real-time method outperforms traditional omnidirectional stereo and lear
ning-based omnidirectional stereo in terms of accuracy and performance.
****
Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework
Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, Hao Li; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4081-4
090
Supervised learning based object detection frameworks demand plenty of laborious
 manual annotations, which may not be practical in real applications. Semi-super
vised object detection (SSOD) can effectively leverage unlabeled data to improve
 the model performance, which is of great significance for the application of ob
ject detection models. In this paper, we revisit SSOD and propose Instant-Teachi
ng, a completely end-to-end and effective SSOD framework, which uses instant pse
udo labeling with extended weak-strong data augmentations for teaching during ea
ch training iteration. To alleviate the confirmation bias problem and improve th
e quality of pseudo annotations, we further propose a co-rectify scheme based on
 Instant-Teaching, denoted as Instant-Teaching*. Extensive experiments on both M
S-COCO and PASCAL VOC datasets substantiate the superiority of our framework. Sp
ecifically, our method surpasses state-of-the-art methods by 4.2 mAP on MS-COCO
when using 2% labeled data. Even with full supervised information of MS-COCO, th
e proposed method still outperforms state-of-the-art methods by about 1.0 mAP. O
n PASCAL VOC, we can achieve more than 5 mAP improvement by applying VOC07 as la
beled data and VOC12 as unlabeled data.
****
Taskology: Utilizing Task Relations at Scale
Yao Lu, Soren Pirk, Jan Dlabal, Anthony Brohan, Ankita Pasad, Zhao Chen, Vincent
 Casser, Anelia Angelova, Ariel Gordon; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8700-8709
Many computer vision tasks address the problem of scene understanding and are na
turally interrelated e.g. object classification, detection, scene segmentation,
depth estimation, etc. We show that we can leverage the inherent relationships a
mong collections of tasks, as they are trained jointly, supervising each other t
hrough their known relationships via consistency losses. Furthermore, explicitly
 utilizing the relationships between tasks allows improving their performance wh
ile dramatically reducing the need for labeled data, and allows training with ad
ditional unsupervised or simulated data. We demonstrate a distributed joint trai
ning algorithm with task-level parallelism, which affords a high degree of async
hronicity and robustness. This allows learning across multiple tasks, or with la
rge amounts of input data, at scale. We demonstrate our framework on subsets of
the following collection of tasks: depth and normal prediction, semantic segment

ation, 3D motion and ego-motion estimation, and object tracking and 3D detection in point clouds. We observe improved performance across these tasks, especially in the low-label regime.
****
Progressive Domain Expansion Network for Single Domain Generalization
Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, Boyang Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 224-233
Single domain generalization is a challenging case of model generalization, where the models are trained on a single domain and tested on other unseen domains. A promising solution is to learn cross-domain invariant representations by expanding the coverage of the training domain. These methods have limited generalization performance gains in practical applications due to the lack of appropriate safety and effectiveness constraints. In this paper, we propose a novel learning framework called progressive domain expansion network (PDEN) for single domain generalization. The domain expansion subnetwork and representation learning subnetwork in PDEN mutually benefit from each other by joint learning. For the domain expansion subnetwork, multiple domains are progressively generated in order to simulate various photometric and geometric transforms in unseen domains. A series of strategies are introduced to guarantee the safety and effectiveness of the expanded domains. For the domain invariant representation learning subnetwork, contrastive learning is introduced to learn the domain invariant representation in which each class is well clustered so that a better decision boundary can be learned to improve it's generalization. Extensive experiments on classification and segmentation have shown that PDEN can achieve up to 15.28% improvement compared with the state-of-the-art single-domain generalization methods.
****
View-Guided Point Cloud Completion
Xuancheng Zhang, Yutong Feng, Siqi Li, Changqing Zou, Hai Wan, Xibin Zhao, Yandong Guo, Yue Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15890-15899
This paper presents a view-guided solution for the task of point cloud completion. Unlike most existing methods directly inferring the missing points using shape priors, we address this task by introducing ViPC (view-guided point cloud completion) that takes the missing crucial global structure information from an extra single-view image. By leveraging a framework which sequentially performs effective cross-modality and cross-level fusions, our method achieves significantly superior results over typical existing solutions on a new large-scale dataset we collect for the view-guided point cloud completion task.
****
Generative Hierarchical Features From Synthesizing Images
Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4432-4442
Generative Adversarial Networks (GANs) have recently advanced image synthesis by learning the underlying distribution of the observed data. However, how the features learned from solving the task of image generation are applicable to other vision tasks remains seldom explored. In this work, we show that learning to synthesize images can bring remarkable hierarchical visual features that are generalizable across a wide range of applications. Specifically, we consider the pre-trained StyleGAN generator as a learned loss function and utilize its layer-wise representation to train a novel hierarchical encoder. The visual feature produced by our encoder, termed as Generative Hierarchical Feature (GH-Feat), has strong transferability to both generative and discriminative tasks, including image editing, image harmonization, image classification, face verification, landmark detection, and layout prediction. Extensive qualitative and quantitative experimental results demonstrate the appealing performance of GH-Feat.
****
Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality
Trisha Mittal, Puneet Mathur, Aniket Bera, Dinesh Manocha; Proceedings of the IE

EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5661-5671

We present Affect2MM, a learning method for time-series emotion prediction for multimedia content. Our goal is to automatically capture the varying emotions depicted by characters in real-life human-centric situations and behaviors. We use the ideas from emotion causation theories to computationally model and determine the emotional state evoked in clips of movies. Affect2MM explicitly models the temporal causality using attention-based methods and Granger causality. We use a variety of components like facial features of actors involved, scene understanding, visual aesthetics, action/situation description, and movie script to obtain an affective-rich representation to understand and perceive the scene. We use an LSTM-based learning model for emotion perception. To evaluate our method, we analyze and compare our performance on three datasets, SENDv1, MovieGraphs, and the LIRIS-ACCEDE dataset, and observe an average of 10-15% increase in the performance over SOTA methods for all three datasets.

****

Black-Box Explanation of Object Detectors via Saliency Maps

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, Kate Saenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11443-11452

We propose D-RISE, a method for generating visual explanations for the predictions of object detectors. Utilizing the proposed similarity metric that accounts for both localization and categorization aspects of object detection allows our method to produce saliency maps that show image areas that most affect the prediction. D-RISE can be considered "black-box" in the software testing sense, as it only needs access to the inputs and outputs of an object detector. Compared to gradient-based methods, D-RISE is more general and agnostic to the particular type of object detector being tested, and does not need knowledge of the inner workings of the model. We show that D-RISE can be easily applied to different object detectors including one-stage detectors such as YOLOv3 and two-stage detectors such as Faster-RCNN. We present a detailed analysis of the generated visual explanations to highlight the utilization of context and possible biases learned by object detectors.

****

Skip-Convolutions for Efficient Video Processing

Amirhossein Habibian, Davide Abati, Taco S. Cohen, Babak Ehteshami Bejnordi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2695-2704

We propose Skip-Convolutions to leverage the large amount of redundancies in video streams and save computations. Each video is represented as a series of changes across frames and network activations, denoted as residuals. We reformulate standard convolution to be efficiently computed on residual frames: each layer is coupled with a binary gate deciding whether a residual is important to the model prediction,e.g. foreground regions, or it can be safely skipped,e.g. background regions. These gates can either be implemented as an efficient network trained jointly with convolution kernels, or can simply skip the residuals based on their magnitude. Gating functions can also incorporate block-wise sparsity structures, as required for efficient implementation on hardware platforms. By replacing all convolutions with Skip-Convolutions in two state-of-the-art architectures, namely EfficientDet and HRNet, we reduce their computational cost consistently by a factor of 3 4x for two different tasks, without any accuracy drop. Extensive comparisons with existing model compression, as well as image and video efficiency methods demonstrate that Skip-Convolutions set a new state-of-the-art by effectively exploiting the temporal redundancies in videos.

****

Looking Into Your Speech: Learning Cross-Modal Affinity for Audio-Visual Speech Separation

Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1336-1345

In this paper, we address the problem of separating individual speech signals fr
om videos using audio-visual neural processing. Most conventional approaches uti
lize frame-wise matching criteria to extract shared information between co-occur
ring audio and video. Thus, their performance heavily depends on the accuracy of
 audio-visual synchronization and the effectiveness of their representations. To
 overcome the frame discontinuity problem between two modalities due to transmis
sion delay mismatch or jitter, we propose a cross-modal affinity network (CaffNe
t) that learns global correspondence as well as locally-varying affinities betwe
en audio and visual streams. Given that the global term provides stability over
a temporal sequence at the utterance-level, this resolves the label permutation
problem characterized by inconsistent assignments. By extending the proposed cro
ss-modal affinity on the complex network, we further improve the separation perf
ormance in the complex spectral domain. Experimental results verify that the pro
posed methods outperform conventional ones on various datasets, demonstrating th
eir advantages in real-world scenarios.
****
GLEAN: Generative Latent Bank for Large-Factor Image Super-Resolution
Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, Chen Change Loy; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2021, pp. 14245-14254
We show that pre-trained Generative Adversarial Networks (GANs), e.g., StyleGAN,
 can be used as a latent bank to improve the restoration quality of large-factor
 image super-resolution (SR). While most existing SR approaches attempt to gener
ate realistic textures through learning with adversarial loss, our method, Gener
ative LatEnt bANk (GLEAN), goes beyond existing practices by directly leveraging
 rich and diverse priors encapsulated in a pre-trained GAN. But unlike prevalent
 GAN inversion methods that require expensive image-specific optimization at run
time, our approach only needs a single forward pass to generate the upscaled ima
ge. GLEAN can be easily incorporated in a simple encoder-bank-decoder architectu
re with multi-resolution skip connections. Switching the bank allows the method
to deal with images from diverse categories, e.g., cat, building, human face, an
d car. Images upscaled by GLEAN shows clear improvements in terms of fidelity an
d texture faithfulness in comparison to existing methods.
****
Soteria: Provable Defense Against Privacy Leakage in Federated Learning From Rep
resentation Perspective
Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, Yiran Chen; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
021, pp. 9311-9319
Federated learning (FL) is a popular distributed learning framework that can red
uce privacy risks by not explicitly sharing private data. However, recent works
have demonstrated that sharing model updates makes FL vulnerable to inference at
tack. In this work, we show our key observation that the data representation lea
kage from gradients is the essential cause of privacy leakage in FL. We also pro
vide an analysis of this observation to explain how the data presentation is lea
ked. Based on this observation, we propose a defense called Soteria against mode
l inversion attack in FL. The key idea of our defense is learning to perturb dat
a representation such that the quality of the reconstructed data is severely deg
raded, while FL performance is maintained. In addition, we derive a certified ro
bustness guarantee to FL and a convergence guarantee to FedAvg, after applying o
ur defense. To evaluate our defense, we conduct experiments on MNIST and CIFAR10
 for defending against the DLG attack and GS attack. Without sacrificing accurac
y, the results demonstrate that our proposed defense can increase the mean squar
ed error between the reconstructed data and the raw data by as much as 160x for
both DLG attack and GS attack, compared with baseline defense methods. Therefore
, the privacy of the FL system is significantly improved. Our code can be found
at https://github.com/jeremy313/Soteria.
****
Deep Occlusion-Aware Instance Segmentation With Overlapping BiLayers
Lei Ke, Yu-Wing Tai, Chi-Keung Tang; Proceedings of the IEEE/CVF Conference on C

omputer Vision and Pattern Recognition (CVPR), 2021, pp. 4019-4028
Segmenting highly-overlapping objects is challenging, because typically no distinction is made between real object contours and occlusion boundaries. Unlike previous two-stage instance segmentation methods, we model image formation as composition of two overlapping layers, and propose Bilayer Convolutional Network (BCNet), where the top GCN layer detects the occluding objects (occluder) and the bottom GCN layer infers partially occluded instance (occludee). The explicit modeling of occlusion relationship with bilayer structure naturally decouples the boundaries of both the occluding and occluded instances, and considers the interaction between them during mask regression. We validate the efficacy of bilayer decoupling on both one-stage and two-stage object detectors with different backbones and network layer choices. Despite its simplicity, extensive experiments on COCO and KINS show that our occlusion-aware BCNet achieves large and consistent performance gain especially for heavy occlusion cases. Code is available at https://github.com/lkeab/BCNet.
****
MonoRec: Semi-Supervised Dense Reconstruction in Dynamic Environments From a Single Moving Camera
Felix Wimbauer, Nan Yang, Lukas von Stumberg, Niclas Zeller, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6112-6122
In this paper, we propose MonoRec, a semi-supervised monocular dense reconstruction architecture that predicts depth maps from a single moving camera in dynamic environments. MonoRec is based on a multi-view stereo setting which encodes the information of multiple consecutive images in a cost volume. To deal with dynamic objects in the scene, we introduce a MaskModule that predicts moving object masks by leveraging the photometric inconsistencies encoded in the cost volumes. Unlike other multi-view stereo methods, MonoRec is able to reconstruct both static and moving objects by leveraging the predicted masks. Furthermore, we present a novel multi-stage training scheme with a semi-supervised loss formulation that does not require LiDAR depth values. We carefully evaluate MonoRec on the KITTI dataset and show that it achieves state-of-the-art performance compared to both multi-view and single-view methods. With the model trained on KITTI, we further demonstrate that MonoRec is able to generalize well to both the Oxford RobotCar dataset and the more challenging TUM-Mono dataset recorded by a handheld camera. Code and related materials are available at https://vision.in.tum.de/research/monorec.
****
DAP: Detection-Aware Pre-Training With Weak Supervision
Yuanyi Zhong, Jianfeng Wang, Lijuan Wang, Jian Peng, Yu-Xiong Wang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4537-4546
This paper presents a detection-aware pre-training (DAP) approach, which leverages only weakly-labeled classification-style datasets (e.g., ImageNet) for pre-training, but is specifically tailored to benefit object detection tasks. In contrast to the widely used image classification-based pre-training (e.g., on ImageNet), which does not include any location-related training tasks, we transform a classification dataset into a detection dataset through a weakly supervised object localization method based on Class Activation Maps to directly pre-train a detector, making the pre-trained model location-aware and capable of predicting bounding boxes. We show that DAP can outperform the traditional classification pre-training in terms of both sample efficiency and convergence speed in downstream detection tasks including VOC and COCO. In particular, DAP boosts the detection accuracy by a large margin when the number of examples in the downstream task is small.
****
Spatial Assembly Networks for Image Representation Learning
Yang Li, Shichao Kan, Jianhe Yuan, Wenming Cao, Zhihai He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13876-13885

It has been long recognized that deep neural networks are sensitive to changes in spatial configurations or scene structures. Image augmentations, such as random translation, cropping, and resizing, can be used to improve the robustness of deep neural networks under spatial transforms. However, changes in object part configurations, spatial layout of object, and scene structures of the images may still result in major changes in the their feature representations generated by the network, creating significant challenges for various visual learning tasks, including representation or metric learning, image classification and retrieval. In this work, we introduce a new learnable module, called spatial assembly network (SAN), to address this important issue. This SAN module examines the input image and performs a learned re-organization and assembly of feature points from different spatial locations conditioned by feature maps from previous network layers so as to maximize the discriminative power of the final feature representation. This differentiable module can be flexibly incorporated into existing network architectures, improving their capabilities in handling spatial variations and structural changes of the image scene. We demonstrate that the proposed SAN module is able to significantly improve the performance of various metric / representation learning, image retrieval and classification tasks, in both supervised and unsupervised learning scenarios.
****
Linguistic Structures As Weak Supervision for Visual Scene Graph Generation
Keren Ye, Adriana Kovashka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8289-8299
Prior work in scene graph generation requires categorical supervision at the level of triplets---subjects and objects, and predicates that relate them, either with or without bounding box information. However, scene graph generation is a holistic task: thus holistic, contextual supervision should intuitively improve performance. In this work, we explore how linguistic structures in captions can benefit scene graph generation. Our method captures the information provided in captions about relations between individual triplets, and context for subjects and objects (e.g. visual properties are mentioned). Captions are a weaker type of supervision than triplets since the alignment between the exhaustive list of human-annotated subjects and objects in triplets, and the nouns in captions, is weak. However, given the large and diverse sources of multimodal data on the web (e.g. blog posts with images and captions), linguistic supervision is more scalable than crowdsourced triplets. We show extensive experimental comparisons against prior methods which leverage instance- and image-level supervision, and ablate our method to show the impact of leveraging phrasal and sequential context, and techniques to improve localization of subjects and objects.
****
SKFAC: Training Neural Networks With Faster Kronecker-Factored Approximate Curvature
Zedong Tang, Fenlong Jiang, Maoguo Gong, Hao Li, Yue Wu, Fan Yu, Zidong Wang, Min Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13479-13487
The bottleneck of computation burden limits the widespread use of the 2nd order optimization algorithms for training deep neural networks. In this paper, we present a computationally efficient approximation for natural gradient descent, named Swift Kronecker-Factored Approximate Curvature (SKFAC), which combines Kronecker factorization and a fast low-rank matrix inversion technique. Our research aims at both fully connected and convolutional layers. For the fully connected layers, by utilizing the low-rank property of Kronecker factors of Fisher information matrix, our method only requires inverting a small matrix to approximate the curvature with desirable accuracy. For convolutional layers, we propose a way with two strategies to save computational efforts without affecting the empirical performance by reducing across the spatial dimension or receptive fields of feature maps. Specifically, we propose two effective dimension reduction methods for this purpose: Spatial Subsampling and Reduce Sum. Experimental results of training several deep neural networks on Cifar-10 and ImageNet-1k datasets demonstrate that SKFAC can capture the main curvature and yield comparative performance t

o K-FAC. The proposed method bridges the wall-clock time gap between the 1st and 2nd order algorithms.
****
Global2Local: Efficient Structure Search for Video Action Segmentation
Shang-Hua Gao, Qi Han, Zhong-Yu Li, Pai Peng, Liang Wang, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16805-16814
Temporal receptive fields of models play an important role in action segmentation. Large receptive fields facilitate the long-term relations among video clips while small receptive fields help capture the local details. Existing methods construct models with hand-designed receptive fields in layers. Can we effectively search for receptive field combinations to replace hand-designed patterns? To answer this question, we propose to find better receptive field combinations through a global-to-local search scheme. Our search scheme exploits both global search to find the coarse combinations and local search to get the refined receptive field combination patterns further. The global search finds possible coarse combinations other than human-designed patterns. On top of the global search, we propose an expectation guided iterative local search scheme to refine combinations effectively. Our global-to-local search can be plugged into existing action segmentation methods to achieve state-of-the-art performance. The source code is publicly available on http://mmcheng.net/g2lsearch.
****
Picasso: A CUDA-Based Library for Deep Learning Over 3D Meshes
Huan Lei, Naveed Akhtar, Ajmal Mian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13854-13864
We present Picasso, a CUDA-based library comprising novel modules for deep learning over complex real-world 3D meshes. Hierarchical neural architectures have proved effective in multi-scale feature extraction which signifies the need for fast mesh decimation. However, existing methods rely on CPU-based implementations to obtain multi-resolution meshes. We design GPU-accelerated mesh decimation to facilitate network resolution reduction efficiently on-the-fly. Pooling and unpooling modules are defined on the vertex clusters gathered during decimation. For feature learning over meshes, Picasso contains three types of novel convolutions namely, facet2vertex, vertex2facet, and facet2facet convolution. Hence, it treats a mesh as a geometric structure comprising vertices and facets, rather than a spacial graph with edges as previous methods do. Picasso also incorporates a fuzzy mechanism in its filters for robustness to mesh sampling (vertex density). It exploits Gaussian mixtures to define fuzzy coefficients for the facet2vertex convolution, and barycentric interpolation to define the coefficients for the remaining two convolutions. In this release, we demonstrate the effectivenss of the proposed modules with competitive segmentation results on S3DIS. The library will be made public through github.
****
DeFlow: Learning Complex Image Degradations From Unpaired Data With Conditional Flows
Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 94-103
The difficulty of obtaining paired data remains a major bottleneck for learning image restoration and enhancement models for real-world applications. Current strategies aim to synthesize realistic training data by modeling noise and degradations that appear in real-world settings. We propose DeFlow, a method for learning stochastic image degradations from unpaired data. Our approach is based on a novel unpaired learning formulation for conditional normalizing flows. We model the degradation process in the latent space of a shared flow encoder-decoder network. This allows us to learn the conditional distribution of a noisy image given the clean input by solely minimizing the negative log-likelihood of the marginal distributions. We validate our DeFlow formulation on the task of joint image restoration and super-resolution. The models trained with the synthetic data generated by DeFlow outperform previous learnable approaches on three recent datase

ts. Code and trained models will be made available at: https://github.com/volflo
w/DeFlow
****
Student-Teacher Learning From Clean Inputs to Noisy Inputs
Guanzhe Hong, Zhiyuan Mao, Xiaojun Lin, Stanley H. Chan; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 120
75-12084
Feature-based student-teacher learning, a training method that encourages the st
udent's hidden features to mimic those of the teacher network, is empirically su
ccessful in transferring the knowledge from a pre-trained teacher network to the
 student network. Furthermore, recent empirical results demonstrate that, the te
acher's features can boost the student network's generalization even when the st
udent's input sample is corrupted by noise. However, there is a lack of theoreti
cal insights into why and when this method of transferring knowledge can be succ
essful between such heterogeneous tasks. We analyze this method theoretically us
ing deep linear networks, and experimentally using nonlinear networks. We identi
fy three vital factors to the success of the method: (1) whether the student is
trained to zero training loss; (2) how knowledgeable the teacher is on the clean
-input problem; (3) how the teacher decomposes its knowledge in its hidden featu
res. Lack of proper control in any of the three factors leads to failure of the
student-teacher learning method.
****
AdvSim: Generating Safety-Critical Scenarios for Self-Driving Vehicles
Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Cas
as, Mengye Ren, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2021, pp. 9909-9918
As self-driving systems become better, simulating scenarios where the autonomy s
tack may fail becomes more important. Traditionally, those scenarios are generat
ed for a few scenes with respect to the planning module that takes ground-truth
actor states as input. This does not scale and cannot identify all possible auto
nomy failures, such as perception failures due to occlusion. In this paper, we p
ropose AdvSim, an adversarial framework to generate safety-critical scenarios fo
r any LiDAR-based autonomy system. Given an initial traffic scenario, AdvSim mod
ifies the actors' trajectories in a physically plausible manner and updates the
LiDAR sensor data to match the perturbed world. Importantly, by simulating direc
tly from sensor data, we obtain adversarial scenarios that are safety-critical f
or the full autonomy stack. Our experiments show that our approach is general an
d can identify thousands of semantically meaningful safety-critical scenarios fo
r a wide range of modern self-driving systems. Furthermore, we show that the rob
ustness and safety of these systems can be further improved by training them wit
h scenarios generated by AdvSim.
****
MoViNets: Mobile Video Networks for Efficient Video Recognition
Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown
, Boqing Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2021, pp. 16020-16030
We present Mobile Video Networks (MoViNets), a family of computation and memory
efficient video networks that can operate on streaming video for online inferenc
e. 3D convolutional neural networks (CNNs) are accurate at video recognition but
 require large computation and memory budgets and do not support online inferenc
e, making them difficult to work on mobile devices. We propose a three-step appr
oach to improve computational efficiency while substantially reducing the peak m
emory usage of 3D CNNs. First, we design a video network search space and employ
 neural architecture search to generate efficient and diverse 3D CNN architectur
es. Second, we introduce the Stream Buffer technique that decouples memory from
video clip duration, allowing 3D CNNs to embed arbitrary-length streaming video
sequences for both training and inference with a small constant memory footprint
. Third, we propose a simple ensembling technique to improve accuracy further wi
thout sacrificing efficiency. These three progressive techniques allow MoViNets
to achieve state-of-the-art accuracy and efficiency on the Kinetics, Moments in

Time, and Charades video action recognition datasets. For instance, MoViNet-A5-Stream achieves the same accuracy as X3D-XL on Kinetics 600 while requiring 80% fewer FLOPs and 65% less memory. Code is available at https://github.com/google-research/movinet.
****
IBRNet: Learning Multi-View Image-Based Rendering
Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4690-4699
We present a method that synthesizes novel views of complex scenes by interpolating a sparse set of nearby views. The core of our method is a network architecture that includes a multilayer perceptron and a ray transformer that estimates radiance and volume density at continuous 5D locations (3D spatial locations and 2D viewing directions), drawing appearance information on the fly from multiple source views. By drawing on source views at render time, our method hearkens back to classic work on image-based rendering (IBR), and allows us to render high-resolution imagery. Unlike neural scene representation work that optimizes per-scene functions for rendering, we learn a generic view interpolation function that generalizes to novel scenes. We render images using classic volume rendering, which is fully differentiable and allows us to train using only multi-view posed images as supervision. Experiments show that our method outperforms recent novel view synthesis methods that also seek to generalize to novel scenes. Further, if fine-tuned on each scene, our method is competitive with state-of-the-art single-scene neural rendering methods.
****
SelfAugment: Automatic Augmentation Policies for Self-Supervised Learning
Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, Kurt Keutzer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2674-2683
A common practice in unsupervised representation learning is to use labeled data to evaluate the quality of the learned representations. This supervised evaluation is then used to guide critical aspects of the training process such as selecting the data augmentation policy. However, guiding an unsupervised training process through supervised evaluations is not possible for real-world data that does not actually contain labels (which may be the case, for example, in privacy sensitive fields such as medical imaging). Therefore, in this work we show that evaluating the learned representations with a self-supervised image rotation task is highly correlated with a standard set of supervised evaluations (rank correlation > 0.94). We establish this correlation across hundreds of augmentation policies, training settings, and network architectures and provide an algorithm (SelfAugment) to automatically and efficiently select augmentation policies without using supervised evaluations. Despite not using any labeled data, the learned augmentation policies perform comparably with augmentation policies that were determined using exhaustive supervised evaluations.
****
Adversarial Invariant Learning
Nanyang Ye, Jingxuan Tang, Huayu Deng, Xiao-Yun Zhou, Qianxiao Li, Zhenguo Li, Guang-Zhong Yang, Zhanxing Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12446-12454
Though machine learning algorithms are able to achieve pattern recognition from the correlation between data and labels, the presence of spurious features in the data decreases the robustness of these learned relationships with respect to varied testing environments. This is known as out-of-distribution (OoD) generalization problem. Recently, invariant risk minimization (IRM) attempts to tackle this issue by penalizing predictions based on the unstable spurious features in the data collected from different environments. However, similar to domain adaptation or domain generalization, a prevalent non-trivial limitation in these works is that the environment information is assigned by human specialists i.e. a priori or determined heuristically. However, an inappropriate group partitioning c

an dramatically deteriorate the OoD generalization and the process is expensive and time-consuming. To deal with this issue, we propose a novel theoretically principled min-max framework to iteratively construct a worst-case splitting, i.e. creating the most challenging environment splittings for the backbone learning paradigm (e.g. IRM) to learn the robust feature representation. We also design a differentiable training strategy to facilitate the feasible gradient-based computation. Numerical experiments show that our algorithmic framework has achieved superior and stable performance in various datasets, such as Colored MNIST and Punctuated Stanford Sentiment Treebank (SST). Furthermore, we also find our algorithm to be robust even to a strong data poisoning attack. To the best of our knowledge, this is one of the first to adopt differentiable environment splitting method to enable stable predictions across environments without environment index information, which achieves the state-of-the-art performance on datasets with strong spurious correlation, such as Colored MNIST.
****
Densely Connected Multi-Dilated Convolutional Networks for Dense Prediction Tasks
Naoya Takahashi, Yuki Mitsufuji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 993-1002
Tasks that involve high-resolution dense prediction require a modeling of both local and global patterns in a large input field. Although the local and global structures often depend on each other and their simultaneous modeling is important, many convolutional neural network (CNN)-based approaches interchange representations in different resolutions only a few times. In this paper, we claim the importance of a dense simultaneous modeling of multiresolution representation and propose a novel CNN architecture called densely connected multidilated DenseNet (D3Net). D3Net involves a novel multidilated convolution that has different dilation factors in a single layer to model different resolutions simultaneously. By combining the multidilated convolution with the DenseNet architecture, D3Net incorporates multiresolution learning with an exponentially growing receptive field in almost all layers, while avoiding the aliasing problem that occurs when we naively incorporate the dilated convolution in DenseNet. Experiments on the image semantic segmentation task using Cityscapes and the audio source separation task using MUSDB18 show that the proposed method has superior performance over state-of-the-art methods.
****
Depth-Conditioned Dynamic Message Propagation for Monocular 3D Object Detection
Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, Li Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 454-463
The objective of this paper is to learn context- and depth-aware feature representation to solve the problem of monocular 3D object detection. We make following contributions: (i) rather than appealing to the complicated pseudo-LiDAR based approach, we propose a depth-conditioned dynamic message propagation (DDMP) network to effectively integrate the multi-scale depth information with the image context; (ii) this is achieved by first adaptively sampling context-aware nodes in the image context and then dynamically predicting hybrid depth-dependent filter weights and affinity matrices for propagating information; (iii) by augmenting a center-aware depth encoding (CDE) task, our method successfully alleviates the inaccurate depth prior; (iv) we thoroughly demonstrate the effectiveness of our proposed approach and show state-of-the-art results among the monocular-based approaches on the KITTI benchmark dataset. Particularly, we rank 1st in the highly competitive KITTI monocular 3D object detection track on the submission day (November 16th, 2020). Code and models are released at https://github.com/fudan-zvg/DDMP
****
S2-BNN: Bridging the Gap Between Self-Supervised Real and 1-Bit Neural Networks via Guided Distribution Calibration
Zhiqiang Shen, Zechun Liu, Jie Qin, Lei Huang, Kwang-Ting Cheng, Marios Savvides; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

ion (CVPR), 2021, pp. 2165-2174
Previous studies dominantly target at self-supervised learning on real-valued ne tworks and have achieved many promising results. However, on the more challengin g binary neural networks (BNNs), this task has not yet been fully explored in th e community. In this paper, we focus on this more difficult scenario: learning n etworks where both weights and activations are binary, meanwhile, without any hu man annotated labels. We observe that the commonly used contrastive objective is not satisfying on BNNs for competitive accuracy, since the backbone network con tains relatively limited capacity and representation ability. Hence instead of d irectly applying existing self-supervised methods, which cause a severe decline in performance, we present a novel guided learning paradigm from real-valued to distill binary networks on the final prediction distribution, to minimize the lo ss and obtain desirable accuracy. Our proposed method can boost the simple contr astive learning baseline by an absolute gain of 5.5 15% on BNNs. We further reve al that it is difficult for BNNs to recover the similar predictive distributions as real-valued models when training without labels. Thus, how to calibrate them is key to address the degradation in performance. Extensive experiments are con ducted on the large-scale ImageNet and downstream datasets. Our method achieves substantial improvement over the simple contrastive learning baseline, and is ev en comparable to many mainstream supervised BNN methods. Code is available at ht tps://github.com/szq0214/S2-BNN.
****
Learning Optical Flow From Still Images
Filippo Aleotti, Matteo Poggi, Stefano Mattoccia; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15201-1521 1
This paper deals with the scarcity of data for training optical flow networks, h ighlighting the limitations of existing sources such as labeled synthetic datase ts or unlabeled real videos. Specifically, we introduce a framework to generate accurate ground-truth optical flow annotations quickly and in large amounts from any readily available single real picture. Given an image, we use an off-the-sh elf monocular depth estimation network to build a plausible point cloud for the observed scene. Then, we virtually move the camera in the reconstructed environm ent with known motion vectors and rotation angles, allowing us to synthesize bot h a novel view and the corresponding optical flow field connecting each pixel in the input image to the one in the new frame. When trained with our data, state- of-the-art optical flow networks achieve superior generalization to unseen real data compared to the same models trained either on annotated synthetic datasets or unlabeled videos, and better specialization if combined with synthetic images .
****
From Shadow Generation To Shadow Removal
Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, Song Wang; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, p p. 4927-4936
Shadow removal is a computer-vision task that aims to restore the image content in shadow regions. While almost all recent shadow-removal methods require shadow -free images for training, in ECCV 2020 Le and Samaras introduces an innovative approach without this requirement by cropping patches with and without shadows f rom shadow images as training samples. However, it is still laborious and time-c onsuming to construct a large amount of such unpaired patches. In this paper, we propose a new G2R-ShadowNet which leverages shadow generation for weakly-superv ised shadow removal by only using a set of shadow images and their corresponding shadow masks for training. The proposed G2R-ShadowNet consists of three sub-net works for shadow generation, shadow removal and refinement, respectively and the y are jointly trained in an end-to-end fashion. In particular, the shadow genera tion sub-net stylises non-shadow regions to be shadow ones, leading to paired da ta for training the shadow-removal sub-net. Extensive experiments on the ISTD da taset and the Video Shadow Removal dataset show that the proposed G2R-ShadowNet achieves competitive performances against the current state of the arts and outp

erforms Le and Samaras' patch-based shadow-removal method.
****

Face Forgery Detection by 3D Decomposition

Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2929-2939

Detecting digital face manipulation has attracted extensive attention due to the potential harms of fake media to the public. However, recent advances have been able to reduce the forgery signals to a low magnitude. Decomposition, which reversibly decomposes the image into several constituent elements, is a promising way to highlight the hidden forgery details. In this paper, we consider a face image as the production of the intervention of the underlying 3D geometry and the lighting environment, and decompose it in a computer graphics view. Specifically, by disentangling the face image into 3D shape, common texture, identity texture, ambient light, and direct light, we find the devil lies in the direct light and the identity texture. Based on this observation, we propose to utilize facial detail, which is the combination of direct light and identity texture, as the clue to detect the subtle forgery patterns. Besides, we highlight the manipulated region with a supervised attention mechanism and introduce a two-stream structure to exploit both face image and facial detail together as a multi-modality task. Extensive experiments indicate the effectiveness of the extra features extracted from the facial detail, and our method achieves the state-of-the-art performance.
****

Unsupervised 3D Shape Completion Through GAN Inversion

Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1768-1777

Most 3D shape completion approaches rely heavily on partial-complete shape pairs and learn in a fully supervised manner. Despite their impressive performances on in-domain data, when generalizing to partial shapes in other forms or real-world partial scans, they often obtain unsatisfactory results due to domain gaps. In contrast to previous fully supervised approaches, in this paper we present ShapeInversion, which introduces Generative Adversarial Network (GAN) inversion to shape completion for the first time. ShapeInversion uses a GAN pre-trained on complete shapes by searching for a latent code that gives a complete shape that best reconstructs the given partial input. In this way, ShapeInversion no longer needs paired training data, and is capable of incorporating the rich prior captured in a well-trained generative model. On the ShapeNet benchmark, the proposed ShapeInversion outperforms the SOTA unsupervised method, and is comparable with supervised methods that are learned using paired data. It also demonstrates remarkable generalization ability, giving robust results for real-world scans and partial inputs of various forms and incompleteness levels. Importantly, ShapeInversion naturally enables a series of additional abilities thanks to the involvement of a pre-trained GAN, such as producing multiple valid complete shapes for an ambiguous partial input, as well as shape manipulation and interpolation.
****

Pseudo 3D Auto-Correlation Network for Real Image Denoising

Xiaowan Hu, Ruijun Ma, Zhihong Liu, Yuanhao Cai, Xiaole Zhao, Yulun Zhang, Haoqian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16175-16184

The extraction of auto-correlation in images has shown great potential in deep learning networks, such as the self-attention mechanism in the channel domain and the self-similarity mechanism in the spatial domain. However, the realization of the above mechanisms mostly requires complicated module stacking and a large number of convolution calculations, which inevitably increases model complexity and memory cost. Therefore, we propose a pseudo 3D auto-correlation network (P3AN) to explore a more efficient way of capturing contextual information in image denoising. On the one hand, P3AN uses fast 1D convolution instead of dense connections to realize criss-cross interaction, which requires less computational reso

urces. On the other hand, the operation does not change the feature size and mak
es it easy to expand. It means that only a simple adaptive fusion is needed to o
btain contextual information that includes both the channel domain and the spati
al domain. Our method built a pseudo 3D auto-correlation attention block through
 1D convolutions and a lightweight 2D structure for more discriminative features
. Extensive experiments have been conducted on three synthetic and four real noi
sy datasets. According to quantitative metrics and visual quality evaluation, th
e P3AN shows great superiority and surpasses state-of-the-art image denoising me
thods.
****
MaxUp: Lightweight Adversarial Training With Data Augmentation Improves Neural N
etwork Training
Chengyue Gong, Tongzheng Ren, Mao Ye, Qiang Liu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2474-2483
We propose MaxUp, an embarrassingly simple, highly effective technique for impro
ving the generalization performance of machine learning models, especially deep
neural networks. The idea is to generate a set of augmented data with some rando
m perturbations or transforms, and minimize the maximum, or worst case loss over
 the augmented data. By doing so, we implicitly introduce a smoothness or robust
ness regularization against the random perturbations, and hence improve the gene
ration performance. For example, in the case of Gaussian perturbation, MaxUp is
asymptotically equivalent to using the gradient norm of the loss as a penalty to
 encourage smoothness. We test MaxUp on a range of tasks, including image classi
fication, language modeling, and adversarial certification, on which MaxUp consi
stently outperforms the existing best baseline methods, without introducing subs
tantial computational overhead. In particular, we improve ImageNet classificatio
n from the accuracy 85.5% without extra data to 85.8%.
****
Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Seman
tic Segmentation
Jungbeom Lee, Eunji Kim, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4071-4080
Weakly supervised semantic segmentation produces a pixel-level localization from
 class labels; but a classifier trained on such labels is likely to restrict its
 focus to a small discriminative region of the target object. AdvCAM is an attri
bution map of an image that is manipulated to increase the classification score
produced by a classifier. This manipulation is realized in an anti-adversarial m
anner, which perturbs the original images along pixel gradients in the opposite
direction from those used in an adversarial attack. It forces regions initially
considered not to be discriminative to become involved in subsequent classificat
ions, and produces attribution maps that successively identify more regions of t
he target object. In addition, we introduce a new regularization procedure that
inhibits both the incorrect attribution of regions unrelated to the target objec
t and excessive concentration of attributions on a small region of that object.
Our method is a post-hoc analysis of a trained classifier, which does not need t
o be altered or retrained. On PASCAL VOC 2012 test images, we achieve mIoUs of 6
8.0 and 76.9 for weakly and semi-supervised semantic segmentation respectively,
which represent a new state-of-the-art.
****
Data-Free Knowledge Distillation for Image Super-Resolution
Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, Yunhe Wang; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2021, pp. 7852-7861
Convolutional network compression methods require training data for achieving ac
ceptable results, but training data is routinely unavailable due to some privacy
 and transmission limitations. Therefore, recent works focus on learning efficie
nt networks without original training data, i.e., data-free model compression. W
herein, most of existing algorithms are developed for image recognition or segme
ntation tasks. In this paper, we study the data-free compression approach for si
ngle image super-resolution (SISR) task which is widely used in mobile phones an

d smart cameras. Specifically, we analyze the relationship between the outputs a nd inputs from the pre-trained network and explore a generator with a series of loss functions for maximally capturing useful information. The generator is then trained for synthesizing training samples which have similar distribution to th at of the original data. To further alleviate the training difficulty of the stu dent network using only the synthetic data, we introduce a progressive distillat ion scheme. Experiments on various datasets and architectures demonstrate that t he proposed method is able to be utilized for effectively learning portable stud ent networks without the original data, e.g., with 0.16dB PSNR drop on Set5 for x2 super resolution. Code will be available at https://github.com/huaweinoah/Dat a-Efficient-Model-Compression.
****

PluckerNet: Learn To Register 3D Line Reconstructions
Liu Liu, Hongdong Li, Haodong Yao, Ruyi Zha; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1842-1852
Aligning two partially-overlapped 3D line reconstructions in Euclidean space is challenging, as we need to simultaneously solve line correspondences and relativ e pose between reconstructions. This paper proposes a neural network based metho d and it has three modules connected in sequence: (i) a Multilayer Perceptron (M LP) based network takes Pluecker representations of lines as inputs, to extract discriminative line-wise features and matchabilities (how likely each line is go ing to have a match), (ii) an Optimal Transport (OT) layer takes two-view line-w ise features and matchabilities as inputs to estimate a 2D joint probability mat rix, with each item describes the matchness of a line pair, and (iii) line pairs with Top-K matching probabilities are fed to a 2-line minimal solver in a RANSA C framework to estimate a six Degree-of-Freedom (6-DoF) rigid transformation. Ex periments on both indoor and outdoor datasets show that registration (rotation a nd translation) precision of our method outperforms baselines significantly.
****

Deep Perceptual Preprocessing for Video Coding
Aaron Chadha, Yiannis Andreopoulos; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2021, pp. 14852-14861
We introduce the concept of rate-aware deep perceptual preprocessing (DPP) for v ideo encoding. DPP makes a single pass over each input frame in order to enhance its visual quality when the video is to be compressed with any codec at any bit rate. The resulting bitstreams can be decoded and displayed at the client side w ithout any post-processing component. DPP comprises a convolutional neural netwo rk that is trained via a composite set of loss functions that incorporates: (i) a perceptual loss based on a trained no reference image quality assessment model , (ii) a reference based fidelity loss expressing L1 and structural similarity a spects, (iii) a motion-based rate loss via block-based transform, quantization a nd entropy estimates that converts the essential components of standard hybrid v ideo encoder designs into a trainable framework. Extensive testing using multipl e quality metrics and AVC, AV1 and VVC encoders shows that DPP+encoder reduces, on average, the bitrate of the corresponding encoder by 11%. This marks the firs t time a server-side neural processing component achieves such savings over the state-of-the-art in video coding.
****

Explaining Classifiers Using Adversarial Perturbations on the Perceptual Ball
Andrew Elliott, Stephen Law, Chris Russell; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10693-10702
We present a simple regularization of adversarial perturbations based upon the p erceptual loss. While the resulting perturbations remain imperceptible to the hu man eye, they differ from existing adversarial perturbations in that they are se mi-sparse alterations that highlight objects and regions of interest while leavi ng the background unaltered. As a semantically meaningful adverse perturbations, it forms a bridge between counterfactual explanations and adversarial perturbat ions in the space of images. We evaluate our approach on several standard explai nability benchmarks, namely, weak localization, insertion deletion, and the poin ting game demonstrating that perceptually regularized counterfactuals are an eff

ective explanation for image-based classifiers.
****

DARCNN: Domain Adaptive Region-Based Convolutional Neural Network for Unsupervised Instance Segmentation in Biomedical Images
Joy Hsu, Wah Chiu, Serena Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1003-1012
In the biomedical domain, there is an abundance of dense, complex data where objects of interest may be challenging to detect or constrained by limits of human knowledge. Labelled domain specific datasets for supervised tasks are often expensive to obtain, and furthermore discovery of novel distinct objects may be desirable for unbiased scientific discovery. Therefore, we propose leveraging the wealth of annotations in benchmark computer vision datasets to conduct unsupervised instance segmentation for diverse biomedical datasets. The key obstacle is thus overcoming the large domain shift from common to biomedical images. We propose a Domain Adaptive Region-based Convolutional Neural Network (DARCNN), that adapts knowledge of object definition from COCO, a large labelled vision dataset, to multiple biomedical datasets. We introduce a domain separation module, a self-supervised representation consistency loss, and an augmented pseudo-labelling stage within DARCNN to effectively perform domain adaptation across such large domain shifts. We showcase DARCNN's performance for unsupervised instance segmentation on numerous biomedical datasets.
****