Dual Super-Resolution Learning for Semantic Segmentation
Li Wang, Dong Li, Yousong Zhu, Lu Tian, Yi Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3774-3783
Current state-of-the-art semantic segmentation methods often apply high-resolution input to attain high performance, which brings large computation budgets and limits their applications on resource-constrained devices. In this paper, we propose a simple and flexible two-stream framework named Dual Super-Resolution Learning (DSRL) to effectively improve the segmentation accuracy without introducing extra computation costs. Specifically, the proposed method consists of three parts: Semantic Segmentation Super-Resolution (SSSR), Single Image Super-Resolution (SISR) and Feature Affinity (FA) module, which can keep high-resolution representations with low-resolution input while simultaneously reducing the model computation complexity. Moreover, it can be easily generalized to other tasks, e.g., human pose estimation. This simple yet effective method leads to strong representations and is evidenced by promising performance on both semantic segmentation and human pose estimation. Specifically, for semantic segmentation on CityScapes, we can achieve \geq2% higher mIoU with similar FLOPs, and keep the performance with 70% FLOPs. For human pose estimation, we can gain \geq2% mAP with the same FLOPs and maintain mAP with 30% fewer FLOPs. Code and models are available at https://github.com/wanglixilinx/DSRL.
****
Deep Unfolding Network for Image Super-Resolution
Kai Zhang, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3217-3226
Learning-based single image super-resolution (SISR) methods are continuously showing superior effectiveness and efficiency over traditional model-based methods, largely due to the end-to-end training. However, different from model-based methods that can handle the SISR problem with different scale factors, blur kernels and noise levels under a unified MAP (maximum a posteriori) framework, learning-based methods generally lack such flexibility. To address this issue, this paper proposes an end-to-end trainable unfolding network which leverages both learningbased methods and model-based methods. Specifically, by unfolding the MAP inference via a half-quadratic splitting algorithm, a fixed number of iterations consisting of alternately solving a data subproblem and a prior subproblem can be obtained. The two subproblems then can be solved with neural modules, resulting in an end-to-end trainable, iterative network. As a result, the proposed network inherits the flexibility of model-based methods to super-resolve blurry, noisy images for different scale factors via a single model, while maintaining the advantages of learning-based methods. Extensive experiments demonstrate the superiority of the proposed deep unfolding network in terms of flexibility, effectiveness and also generalizability.
****
Unsupervised Learning for Intrinsic Image Decomposition From a Single Image
Yunfei Liu, Yu Li, Shaodi You, Feng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3248-3257
Intrinsic image decomposition, which is an essential task in computer vision, aims to infer the reflectance and shading of the scene. It is challenging since it needs to separate one image into two components. To tackle this, conventional methods introduce various priors to constrain the solution, yet with limited performance. Meanwhile, the problem is typically solved by supervised learning methods, which is actually not an ideal solution since obtaining ground truth reflectance and shading for massive general natural scenes is challenging and even impossible. In this paper, we propose a novel unsupervised intrinsic image decomposition framework, which relies on neither labeled training data nor hand-crafted priors. Instead, it directly learns the latent feature of reflectance and shading from unsupervised and uncorrelated data. To enable this, we explore the independence between reflectance and shading, the domain invariant content constraint and the physical constraint. Extensive experiments on both synthetic and real image datasets demonstrate consistently superior performance of the proposed method

.
****
COCAS: A Large-Scale Clothes Changing Person Dataset for Re-Identification
Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3400-3409
Recent years have witnessed great progress in person re-identification (re-id). Several academic benchmarks such as Market1501, CUHK03 and DukeMTMC play important roles to promote the re-id research. To our best knowledge, all the existing benchmarks assume the same person will have the same clothes. While in real-world scenarios, it is very often for a person to change clothes. To address the clothes changing person re-id problem, we construct a novel large-scale re-id benchmark named Clothes Changing Person Set (COCAS), which provides multiple images of the same identity with different clothes. COCAS totally contains 62,382 body images from 5,266 persons. Based on COCAS, we introduce a new person re-id setting for clothes changing problem, where the query includes both a clothes template and a person image taking another clothes. Moreover, we propose a two-branch network named Biometric-Clothes Network (BC-Net) which can effectively integrate biometric and clothes feature for re-id under our setting. Experiments show that it is feasible for clothes changing re-id with clothes templates.
****
Dynamic Convolutions: Exploiting Spatial Sparsity for Faster Inference
Thomas Verelst, Tinne Tuytelaars; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2320-2329
Modern convolutional neural networks apply the same operations on every pixel in an image. However, not all image regions are equally important. To address this inefficiency, we propose a method to dynamically apply convolutions conditioned on the input image. We introduce a residual block where a small gating branch learns which spatial positions should be evaluated. These discrete gating decisions are trained end-to-end using the Gumbel-Softmax trick, in combination with a sparsity criterion. Our experiments on CIFAR, ImageNet, Food-101 and MPII show that our method has better focus on the region of interest and better accuracy than an existing methods, at a lower computational complexity. Moreover, we provide an efficient CUDA implementation of our dynamic convolutions using a gather-scatter approach, achieving a significant improvement in inference speed on MobileNetV2 and ShuffleNetV2. On human pose estimation, a task that is inherently spatially sparse, the processing speed is increased by 60% with no loss in accuracy.
****
Alleviation of Gradient Exploding in GANs: Fake Can Be Real
Song Tao, Jia Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1191-1200
In order to alleviate the notorious mode collapse phenomenon in generative adversarial networks (GANs), we propose a novel training method of GANs in which certain fake samples are considered as real ones during the training process. This strategy can reduce the gradient value that generator receives in the region where gradient exploding happens. We show the process of an unbalanced generation and a vicious circle issue resulted from gradient exploding in practical training, which explains the instability of GANs. We also theoretically prove that gradient exploding can be alleviated by penalizing the difference between discriminator outputs and fake-as-real consideration for very close real and fake samples. Accordingly, Fake-As-Real GAN (FARGAN) is proposed with a more stable training process and a more faithful generated distribution. Experiments on different datasets verify our theoretical analysis.
****
Forward and Backward Information Retention for Accurate Binary Neural Networks
Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, Jingkuan Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2250-2259
Weight and activation binarization is an effective approach to deep neural network compression and can accelerate the inference by leveraging bitwise operations

. Although many binarization methods have improved the accuracy of the model by minimizing the quantization error in forward propagation, there remains a noticeable performance gap between the binarized model and the full-precision one. Our empirical study indicates that the quantization brings information loss in both forward and backward propagation, which is the bottleneck of training accurate binary neural networks. To address these issues, we propose an Information Retention Network (IR-Net) to retain the information that consists in the forward activations and backward gradients. IR-Net mainly relies on two technical contributions: (1) Libra Parameter Binarization (Libra-PB): simultaneously minimizing both quantization error and information loss of parameters by balanced and standardized weights in forward propagation; (2) Error Decay Estimator (EDE): minimizing the information loss of gradients by gradually approximating the sign function in backward propagation, jointly considering the updating ability and accurate gradients. We are the first to investigate both forward and backward processes of binary networks from the unified information perspective, which provides new insight into the mechanism of network binarization. Comprehensive experiments with various network structures on CIFAR-10 and ImageNet datasets manifest that the proposed IR-Net can consistently outperform state-of-the-art quantization methods.

****

Cooling-Shrinking Attack: Blinding the Tracker With Imperceptible Noises
Bin Yan, Dong Wang, Huchuan Lu, Xiaoyun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 990-999
Adversarial attack of CNN aims at deceiving models to misbehave by adding imperceptible perturbations to images. This feature facilitates to understand neural networks deeply and to improve the robustness of deep learning models. Although several works have focused on attacking image classifiers and object detectors, an effective and efficient method for attacking single object trackers of any target in a model-free way remains lacking. In this paper, a cooling-shrinking attack method is proposed to deceive state-of-the-art SiameseRPN-based trackers. An effective and efficient perturbation generator is trained with a carefully designed adversarial loss, which can simultaneously cool hot regions where the target exists on the heatmaps and force the predicted bounding box to shrink, making the tracked target invisible to trackers. Numerous experiments on OTB100, VOT2018, and LaSOT datasets show that our method can effectively fool the state-of-the-art SiameseRPN++ tracker by adding small perturbations to the template or the search regions. Besides, our method has good transferability and is able to deceive other top-performance trackers such as DaSiamRPN, DaSiamRPN-UpdateNet, and DiMP. The source codes are available at https://github.com/MasterBin-IIAU/CSA.

****

Zooming Slow-Mo: Fast and Accurate One-Stage Space-Time Video Super-Resolution
Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P. Allebach, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3370-3379
In this paper, we explore the space-time video super-resolution task, which aims to generate a high-resolution (HR) slow-motion video from a low frame rate (LFR), low-resolution (LR) video. A simple solution is to split it into two sub-tasks: video frame interpolation (VFI) and video super-resolution (VSR). However, temporal interpolation and spatial super-resolution are intra-related in this task. Two-stage methods cannot fully take advantage of the natural property. In addition, state-of-the-art VFI or VSR networks require a large frame-synthesis or reconstruction module for predicting high-quality video frames, which makes the two-stage methods have large model sizes and thus be time-consuming. To overcome the problems, we propose a one-stage space-time video super-resolution framework, which directly synthesizes an HR slow-motion video from an LFR, LR video. Rather than synthesizing missing LR video frames as VFI networks do, we firstly temporally interpolate LR frame features in missing LR video frames capturing local temporal contexts by the proposed feature temporal interpolation network. Then, we propose a deformable ConvLSTM to align and aggregate temporal information simultaneously for better leveraging global temporal contexts. Finally, a deep recon

struction network is adopted to predict HR slow-motion video frames. Extensive e xperiments on benchmark datasets demonstrate that the proposed method not only a chieves better quantitative and qualitative performance but also is more than th ree times faster than recent two-stage state-of-the-art methods, e.g., DAIN+EDVR and DAIN+RBPN.

****

A Hierarchical Graph Network for 3D Object Detection on Point Clouds

Jintai Chen, Biwen Lei, Qingyu Song, Haochao Ying, Danny Z. Chen, Jian Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2020, pp. 392-401

3D object detection on point clouds finds many applications. However, most known point cloud object detection methods did not adequately accommodate the charact eristics (e.g., sparsity) of point clouds, and thus some key semantic informatio n (e.g., shape information) is not well captured. In this paper, we propose a ne w graph convolution (GConv) based hierarchical graph network (HGNet) for 3D obje ct detection, which processes raw point clouds directly to predict 3D bounding b oxes. HGNet effectively captures the relationship of the points and utilizes the multi-level semantics for object detection. Specially, we propose a novel shape -attentive GConv (SA-GConv) to capture the local shape features, by modelling th e relative geometric positions of points to describe object shapes. An SA-GConv based U-shape network captures the multi-level features, which are mapped into a n identical feature space by an improved voting module and then further utilized to generate proposals. Next, a new GConv based Proposal Reasoning Module reason s on the proposals considering the global scene semantics, and the bounding boxe s are then predicted. Consequently, our new framework outperforms state-of-the-a rt methods on two large-scale point cloud datasets, by 4% mean average precisio n (mAP) on SUN RGB-D and by 3% mAP on ScanNet-V2.

****

Online Joint Multi-Metric Adaptation From Frequent Sharing-Subset Mining for Per son Re-Identification

Jiahuan Zhou, Bing Su, Ying Wu; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2020, pp. 2909-2918

Person Re-IDentification (P-RID), as an instance-level recognition problem, stil l remains challenging in computer vision community. Many P-RID works aim to lear n faithful and discriminative features/metrics from offline training data and di rectly use them for the unseen online testing data. However, their performance i s largely limited due to the severe data shifting issue between training and tes ting data. Therefore, we propose an online joint multi-metric adaptation model t o adapt the offline learned P-RID models for the online data by learning a serie s of metrics for all the sharing-subsets. Each sharing-subset is obtained from t he proposed novel frequent sharing-subset mining module and contains a group of testing samples which share strong visual similarity relationships to each other . Unlike existing online P-RID methods, our model simultaneously takes both the sample-specific discriminant and the set-based visual similarity among testing s amples into consideration so that the adapted multiple metrics can refine the di scriminant of all the given testing samples jointly via a multi-kernel late fusi on framework. Our proposed model is generally suitable to any offline learned P- RID baselines for online boosting, the performance improvement by our model is n ot only verified by extensive experiments on several widely-used P-RID benchmark s (CUHK03, Market1501, DukeMTMC-reID and MSMT17) and state-of-the-art P-RID base lines but also guaranteed by the provided in-depth theoretical analyses.

****

Learning to Discriminate Information for Online Action Detection

Hyunjun Eun, Jinyoung Moon, Jongyoul Park, Chanho Jung, Changick Kim; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2020, pp. 809-818

From a streaming video, online action detection aims to identify actions in the present. For this task, previous methods use recurrent networks to model the tem poral sequence of current action frames. However, these methods overlook the fac t that an input image sequence includes background and irrelevant actions as wel

l as the action of interest. For online action detection, in this paper, we prop ose a novel recurrent unit to explicitly discriminate the information relevant t o an ongoing action from others. Our unit, named Information Discrimination Unit (IDU), decides whether to accumulate input information based on its relevance t o the current action. This enables our recurrent network with IDU to learn a mor e discriminative representation for identifying ongoing actions. In experiments on two benchmark datasets, TVSeries and THUMOS-14, the proposed method outperfor ms state-of-the-art methods by a significant margin. Moreover, we demonstrate th e effectiveness of our recurrent unit by conducting comprehensive ablation studi es.
****

Video to Events: Recycling Video Datasets for Event Cameras
Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrio, Davide Scaramuzza; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 3586-3595
Event cameras are novel sensors that output brightness changes in the form of a stream of asynchronous "events" instead of intensity frames. They offer signific ant advantages with respect to conventional cameras: high dynamic range (HDR), h igh temporal resolution, and no motion blur. Recently, novel learning approaches operating on event data have achieved impressive results. Yet, these methods re quire a large amount of event data for training, which is hardly available due t he novelty of event sensors in computer vision research. In this paper, we prese nt a method that addresses these needs by converting any existing video dataset recorded with conventional cameras to synthetic event data. This unlocks the use of a virtually unlimited number of existing video datasets for training network s designed for real event data. We evaluate our method on two relevant vision ta sks, i.e., object recognition and semantic segmentation, and show that models tr ained on synthetic events have several benefits: (i) they generalize well to rea l event data, even in scenarios where standard-camera images are blurry or overe xposed, by inheriting the outstanding properties of event cameras; (ii) they can be used for fine-tuning on real data to improve over state-of-the-art for both classification and semantic segmentation.
****

Bundle Pooling for Polygonal Architecture Segmentation Problem
Huayi Zeng, Kevin Joseph, Adam Vest, Yasutaka Furukawa; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7 50-759
This paper introduces a polygonal architecture segmentation problem, proposes bu ndle-pooling modules for line structure reasoning, and demonstrates a virtual re modeling application that produces production quality results. Given a photograp h of a house with a few vanishing point candidates, we decompose the house into a set of architectural components, each of which is represented as a simple geom etric primitive. A bundle-pooling module pools convolutional features along a bu ndle of line segments (e.g., a family of vanishing lines) and fuses the bundle o f features to determine polygonal boundaries or assign a corresponding vanishing point. Qualitative and quantitative evaluations demonstrate significant improve ments over the existing techniques based on our metric and benchmark dataset. We will share the code and data for further research.
****

Use the Force, Luke! Learning to Predict Physical Forces by Simulating Effects
Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, Abhinav Gupta; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 224-233
When we humans look at a video of human-object interaction, we can not only infe r what is happening but we can even extract actionable information and imitate t hose interactions. On the other hand, current recognition or geometric approache s lack the physicality of action representation. In this paper, we take a step t owards more physical understanding of actions. We address the problem of inferri ng contact points and the physical forces from videos of humans interacting with objects. One of the main challenges in tackling this problem is obtaining groun

d-truth labels for forces. We sidestep this problem by instead using a physics s imulator for supervision. Specifically, we use a simulator to predict effects, a nd enforce that estimated forces must lead to same effect as depicted in the vid eo. Our quantitative and qualitative results show that (a) we can predict meanin gful forces from videos whose effects lead to accurate imitation of the motions observed, (b) by jointly optimizing for contact point and force prediction, we c an improve the performance on both tasks in comparison to independent training, and (c) we can learn a representation from this model that generalizes to novel objects using few shot examples.
****

Articulation-Aware Canonical Surface Mapping
Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, Shubham Tulsiani; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 452-461
We tackle the tasks of: 1) predicting a Canonical Surface Mapping (CSM) that ind icates the mapping from 2D pixels to corresponding points on a canonical templat e shape , and 2) inferring the articulation and pose of the template correspondi ng to the input image. While previous approaches rely on keypoint supervision fo r learning, we present an approach that can learn without such annotations. Our key insight is that these tasks are geometrically related, and we can obtain sup ervisory signal via enforcing consistency among the predictions. We present resu lts across a diverse set of animal object categories, showing that our method ca n learn articulation and CSM prediction from image collections using only foregr ound mask labels for training. We empirically show that allowing articulation he lps learn more accurate CSM prediction, and that enforcing the consistency with predicted CSM is similarly critical for learning meaningful articulation.
****

NeuralScale: Efficient Scaling of Neurons for Resource-Constrained Deep Neural N etworks
Eugene Lee, Chen-Yi Lee; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2020, pp. 1478-1487
Deciding the amount of neurons during the design of a deep neural network to max imize performance is not intuitive. In this work, we attempt to search for the n euron (filter) configuration of a fixed network architecture that maximizes accu racy. Using iterative pruning methods as a proxy, we parametrize the change of t he neuron (filter) number of each layer with respect to the change in parameters , allowing us to efficiently scale an architecture across arbitrary sizes. We al so introduce architecture descent which iteratively refines the parametrized fun ction used for model scaling. The combination of both proposed methods is coined as NeuralScale. To prove the efficiency of NeuralScale in terms of parameters, we show empirical simulations on VGG11, MobileNetV2 and ResNet18 using CIFAR10, CIFAR100 and TinyImageNet as benchmark datasets. Our results show an increase in accuracy of 3.04%, 8.56% and 3.41% for VGG11, MobileNetV2 and ResNet18 on CIFAR 10, CIFAR100 and TinyImageNet respectively under a parameter-constrained setting (output neurons (filters) of default configuration with scaling factor of 0.25) .
****

Transfer Learning From Synthetic to Real-Noise Denoising With Adaptive Instance Normalization
Yoonsik Kim, Jae Woong Soh, Gu Yong Park, Nam Ik Cho; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 348 2-3492
Real-noise denoising is a challenging task because the statistics of real-noise do not follow the normal distribution, and they are also spatially and temporall y changing. In order to cope with various and complex real-noise, we propose a w ell-generalized denoising architecture and a transfer learning scheme. Specifica lly, we adopt an adaptive instance normalization to build a denoiser, which can regularize the feature map and prevent the network from overfitting to the train ing set. We also introduce a transfer learning scheme that transfers knowledge l earned from synthetic-noise data to the real-noise denoiser. From the proposed t

ransfer learning, the synthetic-noise denoiser can learn general features from v
arious synthetic-noise data, and the real-noise denoiser can learn the real-nois
e characteristics from real data. From the experiments, we find that the propose
d denoising method has great generalization ability, such that our network train
ed with synthetic-noise achieves the best performance for Darmstadt Noise Datase
t (DND) among the methods from published papers. We can also see that the propos
ed transfer learning scheme robustly works for real-noise images through the lea
rning with a very small number of labeled data.
****

Variational Context-Deformable ConvNets for Indoor Scene Parsing
Zhitong Xiong, Yuan Yuan, Nianhui Guo, Qi Wang; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3992-4002
Context information is critical for image semantic segmentation. Especially in i
ndoor scenes, the large variation of object scales makes spatial-context an impo
rtant factor for improving the segmentation performance. Thus, in this paper, we
 propose a novel variational context-deformable (VCD) module to learn adaptive r
eceptive-field in a structured fashion. Different from standard ConvNets, which
share fixed-size spatial context for all pixels, the VCD module learns a deforma
ble spatial-context with the guidance of depth information: depth information pr
ovides clues for identifying real local neighborhoods. Specifically, adaptive Ga
ussian kernels are learned with the guidance of multimodal information. By multi
plying the learned Gaussian kernel with standard convolution filters, the VCD mo
dule can aggregate flexible spatial context for each pixel during convolution. T
he main contributions of this work are as follows: 1) a novel VCD module is prop
osed, which exploits learnable Gaussian kernels to enable feature learning with
structured adaptive-context; 2) variational Bayesian probabilistic modeling is i
ntroduced for the training of VCD module, which can make it continuous and more
stable; 3) a perspective-aware guidance module is designed to take advantage of
multi-modal information for RGB-D segmentation. We evaluate the proposed approac
h on three widely-used datasets, and the performance improvement has shown the e
ffectiveness of the proposed method.
****

Augmenting Colonoscopy Using Extended and Directional CycleGAN for Lossy Image T
ranslation
Shawn Mathew, Saad Nadeem, Sruti Kumari, Arie Kaufman; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 46
96-4705
Colorectal cancer screening modalities, such as optical colonoscopy (OC) and vir
tual colonoscopy (VC), are critical for diagnosing and ultimately removing polyp
s (precursors for colon cancer). The non-invasive VC is normally used to inspect
 a 3D reconstructed colon (from computed tomography scans) for polyps and if fou
nd, the OC procedure is performed to physically traverse the colon via endoscope
 and remove these polyps. In this paper, we present a deep learning framework, E
xtended and Directional CycleGAN, for lossy unpaired image-to-image translation
between OC and VC to augment OC video sequences with scale-consistent depth info
rmation from VC and VC with patient-specific textures, color and specular highli
ghts from OC (e.g. for realistic polyp synthesis). Both OC and VC contain struct
ural information, but it is obscured in OC by additional patient-specific textur
e and specular highlights, hence making the translation from OC to VC lossy. The
 existing CycleGAN approaches do not handle lossy transformations. To address th
is shortcoming, we introduce an extended cycle consistency loss, which compares
the geometric structures from OC in the VC domain. This loss removes the need fo
r the CycleGAN to embed OC information in the VC domain. To handle a stronger re
moval of the textures and lighting, a Directional Discriminator is introduced to
 differentiate the direction of translation (by creating paired information for
the discriminator), as opposed to the standard CycleGAN which is direction-agnos
tic. Combining the extended cycle consistency loss and the Directional Discrimin
ator, we show state-of-the-art results on scale-consistent depth inference for p
hantom, textured VC and for real polyp and normal colon video sequences. We also
 present results for realistic pendunculated and flat polyp synthesis from bumps

introduced in 3D VC models.
****
BANet: Bidirectional Aggregation Network With Occlusion Handling for Panoptic Segmentation
Yifeng Chen, Guangchen Lin, Songyuan Li, Omar Bourahla, Yiming Wu, Fangfang Wang, Junyi Feng, Mingliang Xu, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3793-3802
Panoptic segmentation aims to perform instance segmentation for foreground instances and semantic segmentation for background stuff simultaneously. The typical top-down pipeline concentrates on two key issues: 1) how to effectively model the intrinsic interaction between semantic segmentation and instance segmentation, and 2) how to properly handle occlusion for panoptic segmentation. Intuitively, the complementarity between semantic segmentation and instance segmentation can be leveraged to improve the performance. Besides, we notice that using detection/mask scores is insufficient for resolving the occlusion problem. Motivated by these observations, we propose a novel deep panoptic segmentation scheme based on a bidirectional learning pipeline. Moreover, we introduce a plug-and-play occlusion handling algorithm to deal with the occlusion between different object instances. The experimental results on COCO panoptic benchmark validate the effectiveness of our proposed method. Codes will be released soon at https://github.com/Mooonside/BANet.
****
C2FNAS: Coarse-to-Fine Neural Architecture Search for 3D Medical Image Segmentation
Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L. Yuille, Daguang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4126-4135
3D convolution neural networks (CNN) have been proved very successful in parsing organs or tumours in 3D medical images, but it remains sophisticated and time-consuming to choose or design proper 3D networks given different task contexts. Recently, Neural Architecture Search (NAS) is proposed to solve this problem by searching for the best network architecture automatically. However, the inconsistency between search stage and deployment stage often exists in NAS algorithms due to memory constraints and large search space, which could become more serious when applying NAS to some memory and time-consuming tasks, such as 3D medical image segmentation. In this paper, we propose a coarse-to-fine neural architecture search (C2FNAS) to automatically search a 3D segmentation network from scratch without inconsistency on network size or input size. Specifically, we divide the search procedure into two stages: 1) the coarse stage, where we search the macro-level topology of the network, i.e. how each convolution module is connected to other modules; 2) the fine stage, where we search at micro-level for operations in each cell based on previous searched macro-level topology. The coarse-to-fine manner divides the search procedure into two consecutive stages and meanwhile resolves the inconsistency. We evaluate our method on 10 public datasets from Medical Segmentation Decalthon (MSD) challenge, and achieve state-of-the-art performance with the network searched using one dataset, which demonstrates the effectiveness and generalization of our searched models.
****
Seeing the World in a Bag of Chips
Jeong Joon Park, Aleksander Holynski, Steven M. Seitz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1417-1427
We address the dual problems of novel view synthesis and environment reconstruction from hand-held RGBD sensors. Our contributions include 1) modeling highly specular objects, 2) modeling inter-reflections and Fresnel effects, and 3) enabling surface light field reconstruction with the same input needed to reconstruct shape alone. In cases where scene surface has a strong mirror-like material component, we generate highly detailed environment images, revealing room composition, objects, people, buildings, and trees visible through windows. Our approach yields state of the art view synthesis techniques, operates on low dynamic range

imagery, and is robust to geometric and calibration errors.
****
Cascaded Deep Video Deblurring Using Temporal Sharpness Prior
Jinshan Pan, Haoran Bai, Jinhui Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3043-3051
We present a simple and effective deep convolutional neural network (CNN) model for video deblurring. The proposed algorithm mainly consists of optical flow estimation from intermediate latent frames and latent frame restoration steps. It first develops a deep CNN model to estimate optical flow from intermediate latent frames and then restores the latent frames based on the estimated optical flow. To better explore the temporal information from videos, we develop a temporal sharpness prior to constrain the deep CNN model to help the latent frame restoration. We develop an effective cascaded training approach and jointly train the proposed CNN model in an end-to-end manner. We show that exploring the domain knowledge of video deblurring is able to make the deep CNN model more compact and efficient. Extensive experimental results show that the proposed algorithm performs favorably against state-of-the-art methods on the benchmark datasets as well as real-world videos.
****
Reflection Scene Separation From a Single Image
Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Alex C. Kot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2398-2406
For images taken through glass, existing methods focus on the restoration of the background scene by regarding the reflection components as noise. However, the scene reflected by glass surface also contains important information to be recovered, especially for the surveillance or criminal investigations. In this paper, instead of removing reflection components from the mixture image, we aim at recovering reflection scenes from the mixture image. We first propose a strategy to obtain such ground truth and its corresponding input images. Then, we propose a two-stage framework to obtain the visible reflection scene from the mixture image. Specifically, we train the network with a shift-invariant loss which is robust to misalignment between the input and output images. The experimental results show that our proposed method achieves promising results.
****
SmallBigNet: Integrating Core and Contextual Views for Video Classification
Xianhang Li, Yali Wang, Zhipeng Zhou, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1092-1101
Temporal convolution has been widely used for video classification. However, it is performed on spatio-temporal contexts in a limited view, which often weakens its capacity of learning video representation. To alleviate this problem, we propose a concise and novel SmallBig network, with the cooperation of small and big views. For the current time step, the small view branch is used to learn the core semantics, while the big view branch is used to capture the contextual semantics. Unlike traditional temporal convolution, the big view branch can provide the small view branch with the most activated video features from a broader 3D receptive field. Via aggregating such big-view contexts, the small view branch can learn more robust and discriminative spatio-temporal representations for video classification. Furthermore, we propose to share convolution in the small and big view branch, which improves model compactness as well as alleviates overfitting. As a result, our SmallBigNet achieves a comparable model size like 2D CNNs, while boosting accuracy like 3D CNNs. We conduct extensive experiments on the large-scale video benchmarks, e.g., Kinetics400, Something-Something V1 and V2. Our SmallBig network outperforms a number of recent state-of-the-art approaches, in terms of accuracy and/or efficiency. The codes and models will be available on https://github.com/xhl-video/SmallBigNet.
****
From Two Rolling Shutters to One Global Shutter
Cenek Albl, Zuzana Kukelova, Viktor Larsson, Michal Polic, Tomas Pajdla, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa

ttern Recognition (CVPR), 2020, pp. 2505-2513
Most consumer cameras are equipped with electronic rolling shutter, leading to image distortions when the camera moves during image capture. We explore a surprisingly simple camera configuration that makes it possible to undo the rolling shutter distortion: two cameras mounted to have different rolling shutter directions. Such a setup is easy and cheap to build and it possesses the geometric constraints needed to correct rolling shutter distortion using only a sparse set of point correspondences between the two images. We derive equations that describe the underlying geometry for general and special motions and present an efficient method for finding their solutions. Our synthetic and real experiments demonstrate that our approach is able to remove large rolling shutter distortions of all types without relying on any specific scene structure.
****

CvxNet: Learnable Convex Decomposition
Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 31-44
Any solid object can be decomposed into a collection of convex polytopes (in short, convexes). When a small number of convexes are used, such a decomposition can be thought of as a piece-wise approximation of the geometry. This decomposition is fundamental in computer graphics, where it provides one of the most common ways to approximate geometry, for example, in real-time physics simulation. A convex object also has the property of being simultaneously an explicit and implicit representation: one can interpret it explicitly as a mesh derived by computing the vertices of a convex hull, or implicitly as the collection of half-space constraints or support functions. Their implicit representation makes them particularly well suited for neural network training, as they abstract away from the topology of the geometry they need to represent. However, at testing time, convexes can also generate explicit representations - polygonal meshes - which can then be used in any downstream application. We introduce a network architecture to represent a low dimensional family of convexes. This family is automatically derived via an auto-encoding process. We investigate the applications of this architecture including automatic convex decomposition, image to 3D reconstruction, and part-based shape retrieval.
****

RoboTHOR: An Open Simulation-to-Real Embodied AI Platform
Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, Ali Farhadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3164-3174
Visual recognition ecosystems (e.g. ImageNet, Pascal, COCO) have undeniably played a prevailing role in the evolution of modern computer vision. We argue that interactive and embodied visual AI has reached a stage of development similar to visual recognition prior to the advent of these ecosystems. Recently, various synthetic environments have been introduced to facilitate research in embodied AI. Notwithstanding this progress, the crucial question of how well models trained in simulation generalize to reality has remained largely unanswered. The creation of a comparable ecosystem for simulation-to-real embodied AI presents many challenges: (1) the inherently interactive nature of the problem, (2) the need for tight alignments between real and simulated worlds, (3) the difficulty of replicating physical conditions for repeatable experiments, (4) and the associated cost. In this paper, we introduce RoboTHOR to democratize research in interactive and embodied visual AI. RoboTHOR offers a framework of simulated environments paired with physical counterparts to systematically explore and overcome the challenges of simulation-to-real transfer, and a platform where researchers across the globe can remotely test their embodied models in the physical world. As a first benchmark, our experiments show there exists a significant gap between the performance of models trained in simulation when they are tested in both simulations and their carefully constructed physical analogs. We hope that RoboTHOR will sp

ur the next stage of evolution in embodied computer vision.
****

Style Normalization and Restitution for Generalizable Person Re-Identification
Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, Li Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3143-3152
Existing fully-supervised person re-identification (ReID) methods usually suffer from poor generalization capability caused by domain gaps. The key to solving this problem lies in filtering out identity-irrelevant interference and learning domain-invariant person representations. In this paper, we aim to design a generalizable person ReID framework which trains a model on source domains yet is able to generalize/perform well on target domains. To achieve this goal, we propose a simple yet effective Style Normalization and Restitution (SNR) module. Specifically, we filter out style variations (e.g., illumination, color contrast) by Instance Normalization (IN). However, such a process inevitably removes discriminative information. We propose to distill identity-relevant feature from the removed information and restitute it to the network to ensure high discrimination. For better disentanglement, we enforce a dual causal loss constraint in SNR to encourage the separation of identity-relevant features and identity-irrelevant features. Extensive experiments demonstrate the strong generalization capability of our framework. Our models empowered by the SNR modules significantly outperform the state-of-the-art domain generalization approaches on multiple widely-used person ReID benchmarks, and also show superiority on unsupervised domain adaptation.
****

Training Noise-Robust Deep Neural Networks via Meta-Learning
Zhen Wang, Guosheng Hu, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4524-4533
Label noise may significantly degrade the performance of Deep Neural Networks (DNNs). To train noise-robust DNNs, Loss correction (LC) approaches have been introduced. LC approaches assume the noisy labels are corrupted from clean (ground-truth) labels by an unknown noise transition matrix T. The backbone DNNs and T can be trained separately, where T is approximated with prior knowledge. For example, T is constructed by stacking the maximum or mean predic- tions of the samples from each class. In this work, we pro- pose a new loss correction approach, named as Meta Loss Correction (MLC), to directly learn T from data via the meta-learning framework. The MLC is model-agnostic and learns T from data rather than heuristically approximates it using prior knowledge. Extensive evaluations are conducted on computer vision (MNIST, CIFAR-10, CIFAR-100, Cloth- ing1M) and natural language processing (Twitter) datasets. The experimental results show that MLC achieves very com- petitive performance against state-of-the-art approaches.
****

HUMBI: A Large Multiview Dataset of Human Body Expressions
Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, Hyun Soo Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2990-3000
This paper presents a new large multiview dataset called HUMBI for human body expressions with natural clothing. The goal of HUMBI is to facilitate modeling view-specific appearance and geometry of gaze, face, hand, body, and garment from assorted people. 107 synchronized HD cam- eras are used to capture 772 distinctive subjects across gen- der, ethnicity, age, and physical condition. With the mul- tiview image streams, we reconstruct high fidelity body ex- pressions using 3D mesh models, which allows representing view-specific appearance using their canonical atlas. We demonstrate that HUMBI is highly effective in learning and reconstructing a complete human model and is complemen- tary to the existing datasets of human body expressions with limited views and subjects such as MPII-Gaze, Multi-PIE, Human3.6M, and Panoptic Studio datasets.
****

Towards Transferable Targeted Attack
Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, Heng Huang; Proc

eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 641-649
An intriguing property of adversarial examples is their transferability, which suggests that black-box attacks are feasible in real-world applications. Previous works mostly study the transferability on non-targeted setting. However, recent studies show that targeted adversarial examples are more difficult to transfer than non-targeted ones. In this paper, we find there exist two defects that lead to the difficulty in generating transferable examples. First, the magnitude of gradient is decreasing during iterative attack, causing excessive consistency between two successive noises in accumulation of momentum, which is termed as noise curing. Second, it is not enough for targeted adversarial examples to just get close to target class without moving away from true class. To overcome the above problems, we propose a novel targeted attack approach to effectively generate more transferable adversarial examples. Specifically, we first introduce the Poincare distance as the similarity metric to make the magnitude of gradient self-adaptive during iterative attack to alleviate noise curing. Furthermore, we regularize the targeted attack process with metric learning to take adversarial examples away from true label and gain more transferable targeted adversarial examples. Experiments on ImageNet validate the superiority of our approach achieving 8% higher attack success rate over other state-of-the-art methods on average in black-box targeted attack.
****

Supervised Raw Video Denoising With a Benchmark Dataset on Dynamic Scenes
Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, Jingyu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2301-2310
In recent years, the supervised learning strategy for real noisy image denoising has been emerging and has achieved promising results. In contrast, realistic noise removal for raw noisy videos is rarely studied due to the lack of noisy-clean pairs for dynamic scenes. Clean video frames for dynamic scenes cannot be captured with a long-exposure shutter or averaging multi-shots as was done for static images. In this paper, we solve this problem by creating motions for controllable objects, such as toys, and capturing each static moment for multiple times to generate clean video frames. In this way, we construct a dataset with 55 groups of noisy-clean videos with ISO values ranging from 1600 to 25600. To our knowledge, this is the first dynamic video dataset with noisy-clean pairs. Correspondingly, we propose a raw video denoising network (RViDeNet) by exploring the temporal, spatial, and channel correlations of video frames. Since the raw video has Bayer patterns, we pack it into four sub-sequences, i.e RGBG sequences, which are denoised by the proposed RViDeNet separately and finally fused into a clean video. In addition, our network not only outputs a raw denoising result, but also the sRGB result by going through an image signal processing (ISP) module, which enables users to generate the sRGB result with their favourite ISPs. Experimental results demonstrate that our method outperforms state-of-the-art video and raw image denoising algorithms on both indoor and outdoor videos.
****

FDA: Fourier Domain Adaptation for Semantic Segmentation
Yanchao Yang, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4085-4095
We describe a simple method for unsupervised domain adaptation, whereby the discrepancy between the source and target distributions is reduced by swapping the low-frequency spectrum of one with the other. We illustrate the method in semantic segmentation, where densely annotated images are aplenty in one domain (synthetic data), but difficult to obtain in another (real images). Current state-of-the-art methods are complex, some requiring adversarial optimization to render the backbone of a neural network invariant to the discrete domain selection variable. Our method does not require any training to perform the domain alignment, just a simple Fourier Transform and its inverse. Despite its simplicity, it achieves state-of-the-art performance in the current benchmarks, when integrated into a relatively standard semantic segmentation model. Our results indicate that even

simple procedures can discount nuisance variability in the data that more sophisticated methods struggle to learn away.
****

SGAS: Sequential Greedy Architecture Search
Guohao Li, Guocheng Qian, Itzel C. Delgadillo, Matthias Muller, Ali Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1620-1630
Architecture design has become a crucial component of successful deep learning. Recent progress in automatic neural architecture search (NAS) shows a lot of promise. However, discovered architectures often fail to generalize in the final evaluation. Architectures with a higher validation accuracy during the search phase may perform worse in the evaluation. Aiming to alleviate this common issue, we introduce sequential greedy architecture search (SGAS), an efficient method for neural architecture search. By dividing the search procedure into sub-problems, SGAS chooses and prunes candidate operations in a greedy fashion. We apply SGAS to search architectures for Convolutional Neural Networks (CNN) and Graph Convolutional Networks (GCN). Extensive experiments show that SGAS is able to find state-of-the-art architectures for tasks such as image classification, point cloud classification and node classification in protein-protein interaction graphs with minimal computational cost.
****

Instance Segmentation of Biological Images Using Harmonic Embeddings
Victor Kulikov, Victor Lempitsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3843-3851
We present a new instance segmentation approach tailored to biological images, where instances may correspond to individual cells, organisms or plant parts. Unlike instance segmentation for user photographs or road scenes, in biological data object instances may be particularly densely packed, the appearance variation may be particularly low, the processing power may be restricted, while, on the other hand, the variability of sizes of individual instances may be limited. The proposed approach successfully addresses these peculiarities. Our approach describes each object instance using an expectation of a limited number of sine waves with frequencies and phases adjusted to particular object sizes and densities. At train time, a fully-convolutional network is learned to predict the object embeddings at each pixel using a simple pixelwise regression loss, while at test time the instances are recovered using clustering in the embedding space. In the experiments, we show that our approach outperforms previous embedding-based instance segmentation approaches on a number of biological datasets, achieving state-of-the-art on a popular CVPPP benchmark. This excellent performance is combined with computational efficiency that is needed for deployment to domain specialists. The source code of the approach is available at https://github.com/kulikovv/harmonic .
****

Rethinking Zero-Shot Video Classification: End-to-End Training for Realistic Applications
Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, Krzysztof Chalupka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4613-4623
Trained on large datasets, deep learning (DL) can accurately classify videos into hundreds of diverse classes. However, video data is expensive to annotate. Zero-shot learning (ZSL) proposes one solution to this problem. ZSL trains a model once, and generalizes to new tasks whose classes are not present in the training dataset. We propose the first end-to-end algorithm for ZSL in video classification. Our training procedure builds on insights from recent video classification literature and uses a trainable 3D CNN to learn the visual features. This is in contrast to previous video ZSL methods, which use pretrained feature extractors. We also extend the current benchmarking paradigm: Previous techniques aim to make the test task unknown at training time but fall short of this goal. We encourage domain shift across training and test data and disallow tailoring a ZSL model to a specific test dataset. We outperform the state-of-the-art by a wide margi

n. Our code, evaluation procedure and model weights are available online github.com/bbrattoli/ZeroShotVideoClassification.
****

A Multigrid Method for Efficiently Training Video Models
Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, Philipp Krahenbuhl; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 153-162
Training competitive deep video models is an order of magnitude slower than training their counterpart image models. Slow training causes long research cycles, which hinders progress in video understanding research. Following standard practice for training image models, video model training has used a fixed mini-batch shape: a specific number of clips, frames, and spatial size. However, what is the optimal shape? High resolution models perform well, but train slowly. Low resolution models train faster, but are less accurate. Inspired by multigrid methods in numerical optimization, we propose to use variable mini-batch shapes with different spatial-temporal resolutions that are varied according to a schedule. The different shapes arise from resampling the training data on multiple sampling grids. Training is accelerated by scaling up the mini-batch size and learning rate when shrinking the other dimensions. We empirically demonstrate a general and robust grid schedule that yields a significant out-of-the-box training speedup without a loss in accuracy for different models (I3D, non-local, SlowFast), data sets (Kinetics, Something-Something, Charades), and training settings (with and without pre-training, 128 GPUs or 1 GPU). As an illustrative example, the proposed multigrid method trains a ResNet-50 SlowFast network 4.5x faster (wall-clock time, same hardware) while also improving accuracy (+0.8% absolute) on Kinetics-400 compared to baseline training. Code is available online.
****

Attention-Aware Multi-View Stereo
Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, Yawei Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1590-1599
Multi-view stereo is a crucial task in computer vision, that requires accurate and robust photo-consistency among input images for depth estimation. Recent studies have shown that learning-based feature matching and confidence regularization can play a vital role in this task. Nevertheless, how to design good matching confidence volumes as well as effective regularizers for them are still under in-depth study. In this paper, we propose an attention-aware deep neural network "AttMVS" for learning multi-view stereo. In particular, we propose a novel attention-enhanced matching confidence volume, that combines the raw pixel-wise matching confidence from the extracted perceptual features with the contextual information of local scenes, to improve the matching robustness. Furthermore, we develop an attention-guided regularization module, which consists of multilevel ray fusion modules, to hierarchically aggregate and regularize the matching confidence volume into a latent depth probability volume.Experimental results show that our approach achieves the best overall performance on the DTU dataset and the intermediate sequences of Tanks & Temples benchmark over many state-of-the-art MVS algorithms.
****

PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection
Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, Jiashi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 482-490
We propose a single-stage Human-Object Interaction (HOI) detection method that has outperformed all existing methods on HICO-DET dataset at 37 fps on a single Titan XP GPU. It is the first real-time HOI detection method. Conventional HOI detection methods are composed of two stages, i.e., human-object proposals generation, and proposals classification. Their effectiveness and efficiency are limited by the sequential and separate architecture. In this paper, we propose a Parallel Point Detection and Matching (PPDM) HOI detection framework. In PPDM, an HOI

is defined as a point triplet < human point, interaction point, object point>. Human and object points are the center of the detection boxes, and the interaction point is the midpoint of the human and object points. PPDM contains two parallel branches, namely point detection branch and point matching branch. The point detection branch predicts three points. Simultaneously, the point matching branch predicts two displacements from the interaction point to its corresponding human and object points. The human point and the object point originated from the same interaction point are considered as matched pairs. In our novel parallel architecture, the interaction points implicitly provide context and regularization for human and object detection. The isolated detection boxes unlikely to form meaningful HOI triplets are suppressed, which increases the precision of HOI detection. Moreover, the matching between human and object detection boxes is only applied around limited numbers of filtered candidate interaction points, which saves much computational cost. Additionally, we build a new application-oriented database named HOI-A, which serves as a good supplement to the existing datasets.
****

PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, Cynthia Rudin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2437-2445

The primary aim of single-image super-resolution is to construct a high-resolution (HR) image from a corresponding low-resolution (LR) input. In previous approaches, which have generally been supervised, the training objective typically measures a pixel-wise average distance between the super-resolved (SR) and HR images. Optimizing such metrics often leads to blurring, especially in high variance (detailed) regions. We propose an alternative formulation of the super-resolution problem based on creating realistic SR images that downscale correctly. We present a novel super-resolution algorithm addressing this problem, PULSE (Photo Upsampling via Latent Space Exploration), which generates high-resolution, realistic images at resolutions previously unseen in the literature. It accomplishes this in an entirely self-supervised fashion and is not confined to a specific degradation operator used during training, unlike previous methods (which require training on databases of LR-HR image pairs for supervised learning). Instead of starting with the LR image and slowly adding detail, PULSE traverses the high-resolution natural image manifold, searching for images that downscale to the original LR image. This is formalized through the "downscaling loss," which guides exploration through the latent space of a generative model. By leveraging properties of high-dimensional Gaussians, we restrict the search space to guarantee that our outputs are realistic. PULSE thereby generates super-resolved images that both are realistic and downscale correctly. We show extensive experimental results demonstrating the efficacy of our approach in the domain of face super-resolution (also known as face hallucination). Our method outperforms state-of-the-art methods in perceptual quality at higher resolutions and scale factors than previously possible.
****

Discrete Model Compression With Resource Constraint for Deep Neural Networks

Shangqian Gao, Feihu Huang, Jian Pei, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1899-1908

In this paper, we target to address the problem of compression and acceleration of Convolutional Neural Networks (CNNs). Specifically, we propose a novel structural pruning method to obtain a compact CNN with strong discriminative power. To find such networks, we propose an efficient discrete optimization method to directly optimize channel-wise differentiable discrete gate under resource constraint while freezing all the other model parameters. Although directly optimizing discrete variables is a complex non-smooth, non-convex and NP-hard problem, our optimization method can circumvent these difficulties by using the straight-through estimator. Thus, our method is able to ensure that the sub-network discovered within the training process reflects the true sub-network. We further extend th

e discrete gate to its stochastic version in order to thoroughly explore the pot
ential sub-networks. Unlike many previous methods requiring per-layer hyper-para
meters, we only require one hyper-parameter to control FLOPs budget. Moreover, o
ur method is globally discrimination-aware due to the discrete setting. The expe
rimental results on CIFAR-10 and ImageNet show that our method is competitive wi
th state-of-the-art methods.
****

GhostNet: More Features From Cheap Operations
Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, Chang Xu; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2020, pp. 1580-1589
Deploying convolutional neural networks (CNNs) on embedded devices is difficult
due to the limited memory and computation resources. The redundancy in feature m
aps is an important characteristic of those successful CNNs, but has rarely been
investigated in neural architecture design. This paper proposes a novel Ghost m
odule to generate more feature maps from cheap operations. Based on a set of int
rinsic feature maps, we apply a series of linear transformations with cheap cost
to generate many ghost feature maps that could fully reveal information underly
ing intrinsic features. The proposed Ghost module can be taken as a plug-and-pla
y component to upgrade existing convolutional neural networks. Ghost bottlenecks
are designed to stack Ghost modules, and then the lightweight GhostNet can be e
asily established. Experiments conducted on benchmarks demonstrate that the prop
osed Ghost module is an impressive alternative of convolution layers in baseline
models, and our GhostNet can achieve higher recognition performance (e.g. 75.7%
top-1 accuracy) than MobileNetV3 with similar computational cost on the ImageNe
t ILSVRC-2012 classification dataset. Code is available at https://github.com/hu
awei-noah/ghostnet.
****

SDFDiff: Differentiable Rendering of Signed Distance Fields for 3D Shape Optimiz
ation
Yue Jiang, Dantong Ji, Zhizhong Han, Matthias Zwicker; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12
51-1261
We propose SDFDiff, a novel approach for image-based shape optimization using di
fferentiable rendering of 3D shapes represented by signed distance functions (SD
Fs). Compared to other representations, SDFs have the advantage that they can re
present shapes with arbitrary topology, and that they guarantee watertight surfa
ces. We apply our approach to the problem of multi-view 3D reconstruction, where
we achieve high reconstruction quality and can capture complex topology of 3D o
bjects. In addition, we employ a multi-resolution strategy to obtain a robust op
timization algorithm. We further demonstrate that our SDF-based differentiable r
enderer can be integrated with deep learning models, which opens up options for
learning approaches on 3D objects without 3D supervision. In particular, we appl
y our method to single-view 3D reconstruction and achieve state-of-the-art resul
ts.
****

Self2Self With Dropout: Learning Self-Supervised Denoising From Single Image
Yuhui Quan, Mingqin Chen, Tongyao Pang, Hui Ji; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1890-1898
In last few years, supervised deep learning has emerged as one powerful tool for
image denoising, which trains a denoising network over an external dataset of n
oisy/clean image pairs. However, the requirement on a high-quality training data
set limits the broad applicability of the denoising networks. Recently, there ha
ve been a few works that allow training a denoising network on the set of extern
al noisy images only. Taking one step further, this paper proposes a self-superv
ised learning method which only uses the input noisy image itself for training.
In the proposed method, the network is trained with dropout on the pairs of Bern
oulli-sampled instances of the input image, and the result is estimated by avera
ging the predictions generated from multiple instances of the trained model with
dropout. The experiments show that the proposed method not only significantly o

utperforms existing single-image learning or non-learning methods, but also is c
ompetitive to the denoising networks trained on external datasets.
****

A Spatiotemporal Volumetric Interpolation Network for 4D Dynamic Medical Image
Yuyu Guo, Lei Bi, Euijoon Ahn, Dagan Feng, Qian Wang, Jinman Kim; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2020, pp. 4726-4735
Dynamic medical images are often limited in its application due to the large rad
iation doses and longer image scanning and reconstruction times. Existing method
s attempt to reduce the volume samples in the dynamic sequence by interpolating
the volumes between the acquired samples. However, these methods are limited to
either 2D images and/or are unable to support large but periodic variations in t
he functional motion between the image volume samples. In this paper, we present
a spatiotemporal volumetric interpolation network (SVIN) designed for 4D dynami
c medical images. SVIN introduces dual networks: the first is the spatiotemporal
motion network that leverages the 3D convolutional neural network (CNN) for uns
upervised parametric volumetric registration to derive spatiotemporal motion fie
ld from a pair of image volumes; the second is the sequential volumetric interpo
lation network, which uses the derived motion field to interpolate image volumes
, together with a new regression-based module to characterize the periodic motio
n cycles in functional organ structures. We also introduce an adaptive multi-sca
le architecture to capture the volumetric large anatomy motions. Experimental re
sults demonstrated that our SVIN outperformed state-of-the-art temporal medical
interpolation methods and natural video interpolation method that has been exten
ded to support volumetric images. Code is available at [1].
****

Where Am I Looking At? Joint Location and Orientation Estimation by Cross-View M
atching
Yujiao Shi, Xin Yu, Dylan Campbell, Hongdong Li; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4064-407
2
Cross-view geo-localization is the problem of estimating the position and orient
ation (latitude, longitude and azimuth angle) of a camera at ground level given
a large-scale database of geo-tagged aerial (eg., satellite) images. Existing ap
proaches treat the task as a pure location estimation problem by learning discri
minative feature descriptors, but neglect orientation alignment. It is well-reco
gnized that knowing the orientation between ground and aerial images can signifi
cantly reduce matching ambiguity between these two views, especially when the gr
ound-level images have a limited Field of View (FoV) instead of a full field-of-
view panorama. Therefore, we design a Dynamic Similarity Matching network to est
imate cross-view orientation alignment during localization. In particular, we ad
dress the cross-view domain gap by applying a polar transform to the aerial imag
es to approximately align the images up to an unknown azimuth angle. Then, a two
-stream convolutional network is used to learn deep features from the ground and
polar-transformed aerial images. Finally, we obtain the orientation by computin
g the correlation between cross-view features, which also provides a more accura
te measure of feature similarity, improving location recall. Experiments on stan
dard datasets demonstrate that our method significantly improves state-of-the-ar
t performance. Remarkably, we improve the top-1 location recall rate on the CVUS
A dataset by a factor of 1.5x for panoramas with known orientation, by a factor
of 3.3x for panoramas with unknown orientation, and by a factor of 6x for 180-de
gree FoV images with unknown orientation.
****

Towards Large Yet Imperceptible Adversarial Image Perturbations With Perceptual
Color Distance
Zhengyu Zhao, Zhuoran Liu, Martha Larson; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1039-1048
The success of image perturbations that are designed to fool image classifier is
assessed in terms of both adversarial effect and visual imperceptibility. The c
onventional assumption on imperceptibility is that perturbations should strive f

or tight Lp-norm bounds in RGB space. In this work, we drop this assumption by p
ursuing an approach that exploits human color perception, and more specifically,
 minimizing perturbation size with respect to perceptual color distance. Our fir
st approach, Perceptual Color distance C&W (PerC-C&W), extends the widely-used C
&W approach and produces larger RGB perturbations. PerC-C&W is able to maintain
adversarial strength, while contributing to imperceptibility. Our second approac
h, Perceptual Color distance Alternating Loss (PerC-AL), achieves the same outco
me, but does so more efficiently by alternating between the classification loss
and perceptual color difference when updating perturbations. Experimental evalua
tion shows PerC approaches outperform conventional Lp approaches in terms of rob
ustness and transferability, and also demonstrates that the PerC distance can pr
ovide added value on top of existing structure-based methods to creating image p
erturbations.
****

Assessing Image Quality Issues for Real-World Problems
Tai-Yin Chiu, Yinan Zhao, Danna Gurari; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3646-3656
We introduce a new large-scale dataset that links the assessment of image qualit
y issues to two practical vision tasks: image captioning and visual question ans
wering. First, we identify for 39,181 images taken by people who are blind wheth
er each is sufficient quality to recognize the content as well as what quality f
laws are observed from six options. These labels serve as a critical foundation
for us to make the following contributions: (1) a new problem and algorithms for
 deciding whether an image is insufficient quality to recognize the content and
so not captionable, (2) a new problem and algorithms for deciding which of six q
uality flaws an image contains, (3) a new problem and algorithms for deciding wh
ether a visual question is unanswerable due to unrecognizable content versus the
 content of interest being missing from the field of view, and (4) a novel appli
cation of more efficiently creating a large-scale image captioning dataset by au
tomatically deciding whether an image is insufficient quality and so should not
be captioned. We publicly-share our datasets and code to facilitate future exten
sions of this work: https://vizwiz.org.
****

Adaptive Dilated Network With Self-Correction Supervision for Counting
Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, Junjie Yan; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
0, pp. 4594-4603
The counting problem aims to estimate the number of objects in images. Due to la
rge scale variation and labeling deviations, it remains a challenging task. The
static density map supervised learning framework is widely used in existing meth
ods, which uses the Gaussian kernel to generate a density map as the learning ta
rget and utilizes the Euclidean distance to optimize the model. However, the fra
mework is intolerable to the labeling deviations and can not reflect the scale v
ariation. In this paper, we propose an adaptive dilated convolution and a novel
supervised learning framework named self-correction (SC) supervision. In the sup
ervision level, the SC supervision utilizes the outputs of the model to iterativ
ely correct the annotations and employs the SC loss to simultaneously optimize t
he model from both the whole and the individuals. In the feature level, the prop
osed adaptive dilated convolution predicts a continuous value as the specific di
lation rate for each location, which adapts the scale variation better than a di
screte and static dilation rate. Extensive experiments illustrate that our appro
ach has achieved a consistent improvement on four challenging benchmarks. Especi
ally, our approach achieves better performance than the state-of-the-art methods
 on all benchmark datasets.
****

Camouflaged Object Detection
Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, Ling
 Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2020, pp. 2777-2787
We present a comprehensive study on a new task named camouflaged object detectio

n (COD), which aims to identify objects that are "seamlessly" embedded in their surroundings. The high intrinsic similarities between the target object and the background make COD far more challenging than the traditional object detection task. To address this issue, we elaborately collect a novel dataset, called COD10K, which comprises 10,000 images covering camouflaged objects in various natural scenes, over 78 object categories. All the images are densely annotated with category, bounding-box, object-/instance-level, and matting-level labels. This dataset could serve as a catalyst for progressing many vision tasks, e.g., localization, segmentation, and alpha-matting, etc. In addition, we develop a simple but effective framework for COD, termed Search Identification Network (SINet). Without any bells and whistles, SINet outperforms various state-of-the-art object detection baselines on all datasets tested, making it a robust, general framework that can help facilitate future research in COD. Finally, we conduct a large-scale COD study, evaluating 13 cutting-edge models, providing some interesting findings, and showing several potential applications. Our research offers the community an opportunity to explore more in this new field. The code will be available at https://github.com/DengPingFan/SINet/.
****

Why Having 10,000 Parameters in Your Camera Model Is Better Than Twelve
Thomas Schops, Viktor Larsson, Marc Pollefeys, Torsten Sattler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2535-2544
Camera calibration is an essential first step in setting up 3D Computer Vision systems. Commonly used parametric camera models are limited to a few degrees of freedom and thus often do not optimally fit to complex real lens distortion. In contrast, generic camera models allow for very accurate calibration due to their flexibility. Despite this, they have seen little use in practice. In this paper, we argue that this should change. We propose a calibration pipeline for generic models that is fully automated, easy to use, and can act as a drop-in replacement for parametric calibration, with a focus on accuracy. We compare our results to parametric calibrations. Considering stereo depth estimation and camera pose estimation as examples, we show that the calibration error acts as a bias on the results. We thus argue that in contrast to current common practice, generic models should be preferred over parametric ones whenever possible. To facilitate this, we released our calibration pipeline at https://github.com/puzzlepaint/camera_calibration, making both easy-to-use and accurate camera calibration available to everyone.
****

BiDet: An Efficient Binarized Object Detector
Ziwei Wang, Ziyi Wu, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2049-2058
In this paper, we propose a binarized neural network learning method called BiDet for efficient object detection. Conventional network binarization methods directly quantize the weights and activations in one-stage or two-stage detectors with constrained representational capacity, so that the information redundancy in the networks causes numerous false positives and degrades the performance significantly. On the contrary, our BiDet fully utilizes the representational capacity of the binary neural networks for object detection by redundancy removal, through which the detection precision is enhanced with alleviated false positives. Specifically, we generalize the information bottleneck (IB) principle to object detection, where the amount of information in the high-level feature maps is constrained and the mutual information between the feature maps and object detection is maximized. Meanwhile, we learn sparse object priors so that the posteriors are concentrated on informative detection prediction with false positive elimination. Extensive experiments on the PASCAL VOC and COCO datasets show that our method outperforms the state-of-the-art binary neural networks by a sizable margin.
****

Searching for Actions on the Hyperbole
Teng Long, Pascal Mettes, Heng Tao Shen, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp

. 1141-1150
In this paper, we introduce hierarchical action search. Starting from the observation that hierarchies are mostly ignored in the action literature, we retrieve not only individual actions but also relevant and related actions, given an action name or video example as input. We propose a hyperbolic action network, which is centered around a hyperbolic space shared by action hierarchies and videos. Our discriminative hyperbolic embedding projects actions on the shared space while jointly optimizing hypernym-hyponym relations between action pairs and a large margin separation between all actions. The projected actions serve as hyperbolic prototypes that we match with projected video representations. The result is a learned space where videos are positioned in entailment cones formed by different subtrees. To perform search in this space, we start from a query and increasingly enlarge its entailment cone to retrieve hierarchically relevant action videos. Experiments on three action datasets with new hierarchy annotations show the effectiveness of our approach for hierarchical action search by name and by video example, regardless of whether queried actions have been seen or not during training. Our implementation is available at https://github.com/Tenglon/hyperbolic_action
****

SG-NN: Sparse Generative Neural Networks for Self-Supervised Scene Completion of RGB-D Scans
Angela Dai, Christian Diller, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 849-858
We present a novel approach that converts partial and noisy RGB-D scans into high-quality 3D scene reconstructions by inferring unobserved scene geometry. Our approach is fully self-supervised and can hence be trained solely on incomplete, real-world scans. To achieve, self-supervision, we remove frames from a given (incomplete) 3D scan in order to make it even more incomplete; self-supervision is then formulated by correlating the two levels of partialness of the same scan while masking out regions that have never been observed. Through generalization across a large training set, we can then predict 3D scene completions even without seeing any 3D scan of entirely complete geometry. Combined with a new 3D sparse generative convolutional neural network architecture, our method is able to predict highly detailed surfaces in a coarse-to-fine hierarchical fashion that outperform existing state-of-the-art methods by a significant margin in terms of reconstruction quality.
****

Stereoscopic Flash and No-Flash Photography for Shape and Albedo Recovery
Xu Cao, Michael Waechter, Boxin Shi, Ye Gao, Bo Zheng, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3430-3439
We present a minimal imaging setup that harnesses both geometric and photometric approaches for shape and albedo recovery. We adopt a stereo camera and a flashlight to capture a stereo image pair and a flash/no-flash pair. From the stereo image pair, we recover a rough shape that captures low-frequency shape variation without high-frequency details. From the flash/no-flash pair, we derive an image formation model for Lambertian objects under natural lighting, based on which a fine normal map is obtained and fused with the rough shape. Further, we use the flash/no-flash pair for cast shadow detection and albedo canceling, making the shape recovery robust against shadows and albedo variation. We verify the effectiveness of our approach on both synthetic and real-world data.
****

What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation
Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, Xiaowei Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4023-4032
Unsupervised domain adaptation has attracted growing research attention on semantic segmentation. However, 1) most existing models cannot be directly applied into lesions transfer of medical images, due to the diverse appearances of same le

sion among different datasets; 2) equal attention has been paid into all semanti c representations instead of neglecting irrelevant knowledge, which leads to neg ative transfer of untransferable knowledge. To address these challenges, we deve lop a new unsupervised semantic transfer model including two complementary modul es (i.e., T_D and T_F) for endoscopic lesions segmentation, which can alternati vely determine where and how to explore transferable domain-invariant knowledge between labeled source lesions dataset (e.g., gastroscope) and unlabeled target diseases dataset (e.g., enteroscopy). Specifically, T_D focuses on where to tran slate transferable visual information of medical lesions via residual transferab ility-aware bottleneck, while neglecting untransferable visual characterizations . Furthermore, T_F highlights how to augment transferable semantic features of v arious lesions and automatically ignore untransferable representations, which ex plores domain-invariant knowledge and in return improves the performance of T_D. To the end, theoretical analysis and extensive experiments on medical endoscopi c dataset and several non-medical public datasets well demonstrate the superiori ty of our proposed model.
****

Learning to Generate 3D Training Data Through Hybrid Gradient
Dawei Yang, Jia Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 779-789
Synthetic images rendered by graphics engines are a promising source for trainin g deep networks. However, it is challenging to ensure that they can help train a network to perform well on real images, because a graphics-based generation pip eline requires numerous design decisions such as the selection of 3D shapes and the placement of the camera. In this work, we propose a new method that optimize s the generation of 3D training data based on what we call "hybrid gradient". We parametrize the design decisions as a real vector, and combine the approximate gradient and the analytical gradient to obtain the hybrid gradient of the networ k performance with respect to this vector. We evaluate our approach on the task of estimating surface normal, depth or intrinsic decomposition from a single ima ge. Experiments on standard benchmarks show that our approach can outperform the prior state of the art on optimizing the generation of 3D training data, partic ularly in terms of computational efficiency.
****

On Joint Estimation of Pose, Geometry and svBRDF From a Handheld Scanner
Carolin Schmitt, Simon Donne, Gernot Riegler, Vladlen Koltun, Andreas Geiger ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 3493-3503
We propose a novel formulation for joint recovery of camera pose, object geometr y and spatially-varying BRDF. The input to our approach is a sequence of RGB-D i mages captured by a mobile, hand-held scanner that actively illuminates the scen e with point light sources. Compared to previous works that jointly estimate geo metry and materials from a hand-held scanner, we formulate this problem using a single objective function that can be minimized using off-the-shelf gradient-bas ed solvers. By integrating material clustering as a differentiable operation int o the optimization process, we avoid pre-processing heuristics and demonstrate t hat our model is able to determine the correct number of specular materials inde pendently. We provide a study on the importance of each component in our formula tion and on the requirements of the initial geometry. We show that optimizing ov er the poses is crucial for accurately recovering fine details and show that our approach naturally results in a semantically meaningful material segmentation.
****

Synchronizing Probability Measures on Rotations via Optimal Transport
Tolga Birdal, Michael Arbel, Umut Simsekli, Leonidas J. Guibas; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 0, pp. 1569-1579
We introduce a new paradigm, `measure synchronization', for synchronizing graphs with measure-valued edges. We formulate this problem as maximization of the cyc le-consistency in the space of probability measures over relative rotations. In particular, we aim at estimating marginal distributions of absolute orientations

by synchronizing the `conditional' ones, which are defined on the Riemannian manifold of quaternions. Such graph optimization on distributions-on-manifolds enables a natural treatment of multimodal hypotheses, ambiguities and uncertainties arising in many computer vision applications such as SLAM, SfM, and object pose estimation. We first formally define the problem as a generalization of the classical rotation graph synchronization, where in our case the vertices denote probability measures over rotations. We then measure the quality of the synchronization by using Sinkhorn divergences, which reduces to other popular metrics such as Wasserstein distance or the maximum mean discrepancy as limit cases. We propose a nonparametric Riemannian particle optimization approach to solve the problem. Even though the problem is non-convex, by drawing a connection to the recently proposed sparse optimization methods, we show that the proposed algorithm converges to the global optimum in a special case of the problem under certain conditions. Our qualitative and quantitative experiments show the validity of our approach and we bring in new perspectives to the study of synchronization.

****

Camera Trace Erasing

Chang Chen, Zhiwei Xiong, Xiaoming Liu, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2950-2959

Camera trace is a unique noise produced in digital imaging process. Most existing forensic methods analyze camera trace to identify image origins. In this paper, we address a new low-level vision problem, camera trace erasing, to reveal the weakness of trace-based forensic methods. A comprehensive investigation on existing anti-forensic methods reveals that it is non-trivial to effectively erase camera trace while avoiding the destruction of content signal. To reconcile these two demands, we propose Siamese Trace Erasing (SiamTE), in which a novel hybrid loss is designed on the basis of Siamese architecture for network training. Specifically, we propose embedded similarity, truncated fidelity, and cross identity to form the hybrid loss. Compared with existing anti-forensic methods, SiamTE has a clear advantage for camera trace erasing, which is demonstrated in three representative tasks.

****

Robust 3D Self-Portraits in Seconds

Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1344-1353

In this paper, we propose an efficient method for robust 3D self-portraits using a single RGBD camera. Benefiting from the proposed PIFusion and lightweight bundle adjustment algorithm, our method can generate detailed 3D self-portraits in seconds and shows the ability to handle subjects wearing extremely loose clothes. To achieve highly efficient and robust reconstruction, we propose PIFusion, which combines learning-based 3D recovery with volumetric non-rigid fusion to generate accurate sparse partial scans of the subject. Moreover, a non-rigid volumetric deformation method is proposed to continuously refine the learned shape prior. Finally, a lightweight bundle adjustment algorithm is proposed to guarantee that all the partial scans can not only "loop" with each other but also remain consistent with the selected live key observations. The results and experiments show that the proposed method achieves more robust and efficient 3D self-portraits compared with state-of-the-art methods.

****

Instance Shadow Detection

Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1880-1889

Instance shadow detection is a brand new problem, aiming to find shadow instances paired with object instances. To approach it, we first prepare a new dataset called SOBA, named after Shadow-OBject Association, with 3,623 pairs of shadow and object instances in 1,000 photos, each with individual labeled masks. Second, we design LISA, named after Light-guided Instance Shadow-object Association, an

end-to-end framework to automatically predict the shadow and object instances, t
ogether with the shadow-object associations and light direction. Then, we pair u
p the predicted shadow and object instances, and match them with the predicted s
hadow-object associations to generate the final results. In our evaluations, we
formulate a new metric named the shadow-object average precision to measure the
performance of our results. Further, we conducted various experiments and demons
trate our method's applicability on light direction estimation and photo editing
.
****

MemNAS: Memory-Efficient Neural Architecture Search With Grow-Trim Learning
Peiye Liu, Bo Wu, Huadong Ma, Mingoo Seok; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2108-2116
Recent studies on automatic neural architecture search techniques have demonstra
ted significant performance, competitive to or even better than hand-crafted neu
ral architectures. However, most of the existing search approaches tend to use r
esidual structures and a concatenation connection between shallow and deep featu
res. A resulted neural network model, therefore, is non-trivial for resource-con
straint devices to execute since such a model requires large memory to store net
work parameters and intermediate feature maps along with excessive computing com
plexity. To address this challenge, we propose MemNAS, a novel growing and trimm
ing based neural architecture search framework that optimizes not only performan
ce but also memory requirement of an inference network. Specifically, in the sea
rch process, we consider running memory use, including network parameters and th
e essential intermediate feature maps memory requirement, as an optimization obj
ective along with performance. Besides, to improve the accuracy of the search, w
e extract the correlation information among multiple candidate architectures to
rank them and then choose the candidates with desired performance and memory eff
iciency. On the ImageNet classification task, our MemNAS achieves 75.4% accuracy
, 0.7% higher than MobileNetV2 with 42.1% less memory requirement. Additional ex
periments confirm that the proposed MemNAS can perform well across the different
 targets of the trade-off between accuracy and memory consumption.
****

Deep Distance Transform for Tubular Structure Segmentation in CT Scans
Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K
. Fishman, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2020, pp. 3833-3842
Tubular structure segmentation in medical images, e.g., segmenting vessels in CT
 scans, serves as a vital step in the use of computers to aid in screening early
 stages of related diseases. But automatic tubular structure segmentation in CT
scans is a challenging problem, due to issues such as poor contrast, noise and c
omplicated background. A tubular structure usually has a cylinder-like shape whi
ch can be well represented by its skeleton and cross-sectional radii (scales). I
nspired by this, we propose a geometry-aware tubular structure segmentation meth
od, Deep Distance Transform (DDT), which combines intuitions from the classical
distance transform for skeletonization and modern deep segmentation networks. DD
T first learns a multi-task network to predict a segmentation mask for a tubular
 structure and a distance map. Each value in the map represents the distance fro
m each tubular structure voxel to the tubular structure surface. Then the segmen
tation mask is refined by leveraging the shape prior reconstructed from the dist
ance map. We apply our DDT on six medical image datasets. Results show that (1)
DDT can boost tubular structure segmentation performance significantly (e.g., ov
er 13% DSC improvement for pancreatic duct segmentation), and (2) DDT additional
ly provides a geometrical measurement for a tubular structure, which is importan
t for clinical diagnosis (e.g., the cross-sectional scale of a pancreatic duct c
an be an indicator for pancreatic cancer).
****

FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding
Dian Shao, Yue Zhao, Bo Dai, Dahua Lin; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2616-2625
On public benchmarks, current action recognition techniques have achieved great

success. However, when used in real-world applications, e.g. sport analysis, whi ch requires the capability of parsing an activity into phases and differentiatin g between subtly different actions, their performances remain far from being sat isfactory. To take action recognition to a new level, we develop FineGym, a new dataset built on top of gymnasium videos. Compared to existing action recognitio n datasets, FineGym is distinguished in richness, quality, and diversity. In par ticular, it provides temporal annotations at both action and sub-action levels w ith a three-level semantic hierarchy. For example, a "balance beam" activity wil l be annotated as a sequence of elementary sub-actions derived from five sets: " leap-jump-hop", "beam-turns", "flight-salto", "flight-handspring", and "dismount ", where the sub-action in each set will be further annotated with finely define d class labels. This new level of granularity presents significant challenges fo r action recognition, e.g. how to parse the temporal structures from a coherent action, and how to distinguish between subtly different action classes. We syste matically investigates different methods on this dataset and obtains a number of interesting findings. We hope this dataset could advance research towards actio n understanding.

****

What Does Plate Glass Reveal About Camera Calibration?

Qian Zheng, Jinnan Chen, Zhan Lu, Boxin Shi, Xudong Jiang, Kim-Hui Yap, Li ng-Yu Duan, Alex C. Kot; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2020, pp. 3022-3032

This paper aims to calibrate the orientation of glass and the field of view of t he camera from a single reflection-contaminated image. We show how a reflective amplitude coefficient map can be used as a calibration cue. Different from exist ing methods, the proposed solution is free from image contents. To reduce the im pact of a noisy calibration cue estimated from a reflection-contaminated image, we propose two strategies: an optimization-based method that imposes part of tho ugh reliable entries on the map and a learning-based method that fully exploits all entries. We collect a dataset containing 320 samples as well as their camera parameters for evaluation. We demonstrate that our method not only facilitates a general single image camera calibration method that leverages image contents b ut also contributes to improving the performance of single image reflection remo val. Furthermore, we show our byproduct output helps alleviate the ill-posed pro blem of estimating the panorama from a single image.

****

One Man's Trash Is Another Man's Treasure: Resisting Adversarial Examples by Adv ersarial Examples

Chang Xiao, Changxi Zheng; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2020, pp. 412-421

Modern image classification systems are often built on deep neural networks, whi ch suffer from adversarial examples--images with deliberately crafted, impercept ible noise to mislead the network's classification. To defend against adversaria l examples, a plausible idea is to obfuscate the network's gradient with respect to the input image. This general idea has inspired a long line of defense metho ds. Yet, almost all of them have proven vulnerable. We revisit this seemingly fl awed idea from a radically different perspective. We embrace the omnipresence of adversarial examples and the numerical procedure of crafting them, and turn thi s harmful attacking process into a useful defense mechanism. Our defense method is conceptually simple: before feeding an input image for classification, transf orm it by finding an adversarial example on a pre-trained external model. We eva luate our method against a wide range of possible attacks. On both CIFAR-10 and Tiny ImageNet datasets, our method is significantly more robust than state-of-th e-art methods. Particularly, in comparison to adversarial training, our method o ffers lower training cost as well as stronger robustness.

****

Image Processing Using Multi-Code GAN Prior

Jinjin Gu, Yujun Shen, Bolei Zhou; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2020, pp. 3012-3021

Despite the success of Generative Adversarial Networks (GANs) in image synthesis

, applying trained GAN models to real image processing remains challenging. Previous methods typically invert a target image back to the latent space either by back-propagation or by learning an additional encoder. However, the reconstructions from both of the methods are far from ideal. In this work, we propose a novel approach, called mGANprior, to incorporate the well-trained GANs as effective prior to a variety of image processing tasks. In particular, we employ multiple latent codes to generate multiple feature maps at some intermediate layer of the generator, then compose them with adaptive channel importance to recover the input image. Such an over-parameterization of the latent space significantly improves the image reconstruction quality, outperforming existing competitors. The resulting high-fidelity image reconstruction enables the trained GAN models as prior to many real-world applications, such as image colorization, super-resolution, image inpainting, and semantic manipulation. We further analyze the properties of the layer-wise representation learned by GAN models and shed light on what knowledge each layer is capable of representing.

****

ColorFool: Semantic Adversarial Colorization

Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, Andrea Cavallaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1151-1160

Adversarial attacks that generate small Lp norm perturbations to mislead classifiers have limited success in black-box settings and with unseen classifiers. These attacks are also not robust to defenses that use denoising filters and to adversarial training procedures. Instead, adversarial attacks that generate unrestricted perturbations are more robust to defenses, are generally more successful in black-box settings and are more transferable to unseen classifiers. However, unrestricted perturbations may be noticeable to humans. In this paper, we propose a content-based black-box adversarial attack that generates unrestricted perturbations by exploiting image semantics to selectively modify colors within chosen ranges that are perceived as natural by humans. We show that the proposed approach, ColorFool, outperforms in terms of success rate, robustness to defense frameworks and transferability, five state-of-the-art adversarial attacks on two different tasks, scene and object classification, when attacking three state-of-the-art deep neural networks using three standard datasets. The source code is available at https://github.com/smartcameras/ColorFool.

****

Bi3D: Stereo Depth Estimation via Binary Classifications

Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, Orazio Gallo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1600-1608

Stereo-based depth estimation is a cornerstone of computer vision, with state-of-the-art methods delivering accurate results in real time. For several applications such as autonomous navigation, however, it may be useful to trade accuracy for lower latency. We present Bi3D, a method that estimates depth via a series of binary classifications. Rather than testing if objects are at a particular depth D, as existing stereo methods do, it classifies them as being closer or farther than D. This property offers a powerful mechanism to balance accuracy and latency. Given a strict time budget, Bi3D can detect objects closer than a given distance in as little as a few milliseconds, or estimate depth with arbitrarily coarse quantization, with complexity linear with the number of quantization levels. Bi3D can also use the allotted quantization levels to get continuous depth, but in a specific depth range. For standard stereo (i.e., continuous depth on the whole range), our method is close to or on par with state-of-the-art, finely tuned stereo methods.

****

D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry

Nan Yang, Lukas von Stumberg, Rui Wang, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1281-1292

We propose D3VO as a novel framework for monocular visual odometry that exploits

deep networks on three levels -- deep depth, pose and uncertainty estimation. We first propose a novel self-supervised monocular depth estimation network trained on stereo videos without any external supervision. In particular, it aligns the training image pairs into similar lighting condition with predictive brightness transformation parameters. Besides, we model the photometric uncertainties of pixels on the input images, which improves the depth estimation accuracy and provides a learned weighting function for the photometric residuals in direct (feature-less) visual odometry. Evaluation results show that the proposed network outperforms state-of-the-art self-supervised depth estimation networks. D3VO tightly incorporates the predicted depth, pose and uncertainty into a direct visual odometry method to boost both the front-end tracking as well as the back-end non-linear optimization. We evaluate D3VO in terms of monocular visual odometry on both the KITTI odometry benchmark and the EuRoC MAV dataset. The results show that D3VO outperforms state-of-the-art traditional monocular VO methods by a large margin. It also achieves comparable results to state-of-the-art stereo/LiDAR odometry on KITTI and to the state-of-the-art visual-inertial odometry on EuRoC MAV, while using only a single camera.

****

Fantastic Answers and Where to Find Them: Immersive Question-Directed Visual Attention

Ming Jiang, Shi Chen, Jinhui Yang, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2980-2989

While most visual attention studies focus on bottom-up attention with restricted field-of-view, real-life situations are filled with embodied vision tasks. The role of attention is more significant in the latter due to the information overload, and attention to the most important regions is critical to the success of tasks. The effects of visual attention on task performance in this context have also been widely ignored. This research addresses a number of challenges to bridge this research gap, on both the data and model aspects. Specifically, we introduce the first dataset of top-down attention in immersive scenes. The Immersive Question-directed Visual Attention (IQVA) dataset features visual attention and corresponding task performance (i.e., answer correctness). It consists of 975 questions and answers collected from people viewing 360deg videos in a head-mounted display. Analyses of the data demonstrate a significant correlation between people's task performance and their eye movements, suggesting the role of attention in task performance. With that, a neural network is developed to encode the differences of correct and incorrect attention and jointly predict the two. The proposed attention model for the first time takes into account answer correctness, whose outputs naturally distinguish important regions from distractions. This study with new data and features may enable new tasks that leverage attention and answer correctness, and inspire new research that reveals the process behind decision making in performing various tasks.

****

Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction

Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 214-223

We propose novel dynamic multiscale graph neural networks (DMGNN) to predict 3D skeleton-based human motions. The core idea of DMGNN is to use a multiscale graph to comprehensively model the internal relations of a human body for motion feature learning. This multiscale graph is adaptive during training and dynamic across network layers. Based on this graph, we propose a multiscale graph computational unit (MGCU) to extract features at individual scales and fuse features across scales. The entire model is action-category-agnostic and follows an encoder-decoder framework. The encoder consists of a sequence of MGCUs to learn motion features. The decoder uses a proposed graph-based gate recurrent unit to generate future poses. Extensive experiments show that the proposed DMGNN outperforms state-of-the-art methods in both short and long-term predictions on the datasets of Human 3.6M and CMU Mocap. We further investigate the learned multiscale graphs

for the interpretability. The codes could be downloaded from https://github.com/limaosen0/DMGNN.
****

Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes From a Single Image
Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, Jian Jun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 55-64
Semantic reconstruction of indoor scenes refers to both scene understanding and object reconstruction. Existing works either address one part of this problem or focus on independent objects. In this paper, we bridge the gap between understanding and reconstruction, and propose an end-to-end solution to jointly reconstruct room layout, object bounding boxes and meshes from a single image. Instead of separately resolving scene understanding and object reconstruction, our method builds upon a holistic scene context and proposes a coarse-to-fine hierarchy with three components: 1. room layout with camera pose; 2. 3D object bounding boxes; 3. object meshes. We argue that understanding the context of each component can assist the task of parsing the others, which enables joint understanding and reconstruction. The experiments on the SUN RGB-D and Pix3D datasets demonstrate that our method consistently outperforms existing methods in indoor layout estimation, 3D object detection and mesh reconstruction.
****

GPS-Net: Graph Property Sensing Network for Scene Graph Generation
Xin Lin, Changxing Ding, Jinquan Zeng, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3746-3753
Scene graph generation (SGG) aims to detect objects in an image along with their pairwise relationships. There are three key properties of scene graph that have been underexplored in recent works: namely, the edge direction information, the difference in priority between nodes, and the long-tailed distribution of relationships. Accordingly, in this paper, we propose a Graph Property Sensing Network (GPS-Net) that fully explores these three properties for SGG. First, we propose a novel message passing module that augments the node feature with node-specific contextual information and encodes the edge direction information via a tri-linear model. Second, we introduce a node priority sensitive loss to reflect the difference in priority between nodes during training. This is achieved by designing a mapping function that adjusts the focusing parameter in the focal loss. Third, since the frequency of relationships is affected by the long-tailed distribution problem, we mitigate this issue by first softening the distribution and then enabling it to be adjusted for each subject-object pair according to their visual appearance. Systematic experiments demonstrate the effectiveness of the proposed techniques. Moreover, GPS-Net achieves state-of-the-art performance on three popular databases: VG, OI, and VRD by significant gains under various settings and metrics. The code and models are available at https://github.com/taksau/GPS-Net.
****

Through the Looking Glass: Neural 3D Reconstruction of Transparent Shapes
Zhengqin Li, Yu-Ying Yeh, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1262-1271
Recovering the 3D shape of transparent objects using a small number of unconstrained natural images is an ill-posed problem. Complex light paths induced by refraction and reflection have prevented both traditional and deep multiview stereo from solving this challenge. We propose a physically-based network to recover 3D shape of transparent objects using a few images acquired with a mobile phone camera, under a known but arbitrary environment map. Our novel contributions include a normal representation that enables the network to model complex light transport through local computation, a rendering layer that models refractions and reflections, a cost volume specifically designed for normal refinement of transparent shapes and a feature mapping based on predicted normals for 3D point cloud reconstruction. We render a synthetic dataset to encourage the model to learn ref

ractive light transport across different views. Our experiments show successful recovery of high-quality 3D geometry for complex transparent shapes using as few as 5-12 natural images. Code and data will be publicly released.Recovering the 3D shape of transparent objects using a small number of unconstrained natural images is an ill-posed problem. Complex light paths induced by refraction and reflection have prevented both traditional and deep multiview stereo from solving this challenge. We propose a physically-based network to recover 3D shape of transparent objects using a few images acquired with a mobile phone camera, under a known but arbitrary environment map. Our novel contributions include a normal representation that enables the network to model complex light transport through local computation, a rendering layer that models refractions and reflections, a cost volume specifically designed for normal refinement of transparent shapes and a feature mapping based on predicted normals for 3D point cloud reconstruction. We render a synthetic dataset to encourage the model to learn refractive light transport across different views. Our experiments show successful recovery of high-quality 3D geometry for complex transparent shapes using as few as 5-12 natural images. Code and data will be publicly released.

****

Recursive Social Behavior Graph for Trajectory Prediction

Jianhua Sun, Qinhong Jiang, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 660-669

Social interaction is an important topic in human trajectory prediction to generate plausible paths. In this paper, we present a novel insight of group-based social interaction model to explore relationships among pedestrians. We recursively extract social representations supervised by group-based annotations and formulate them into a social behavior graph, called Recursive Social Behavior Graph. Our recursive mechanism explores the representation power largely. Graph Convolutional Neural Network then is used to propagate social interaction information in such a graph. With the guidance of Recursive Social Behavior Graph, we surpass state-of-the-art methods on ETH and UCY dataset for 11.1% in ADE and 10.8% in FDE in average, and successfully predict complex social behaviors.

****

Attention Scaling for Crowd Counting

Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, Yanwei Pang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4706-4715

Convolutional Neural Network (CNN) based methods generally take crowd counting as a regression task by outputting crowd densities. They learn the mapping between image contents and crowd density distributions. Though having achieved promising results, these data-driven counting networks are prone to overestimate or underestimate people counts of regions with different density patterns, which degrades the whole count accuracy. To overcome this problem, we propose an approach to alleviate the counting performance differences in different regions. Specifically, our approach consists of two networks named Density Attention Network (DANet) and Attention Scaling Network (ASNet). DANet provides ASNet with attention masks related to regions of different density levels. ASNet first generates density maps and scaling factors and then multiplies them by attention masks to output separate attention-based density maps. These density maps are summed to give the final density map. The attention scaling factors help attenuate the estimation errors in different regions. Furthermore, we present a novel Adaptive Pyramid Loss (APLoss) to hierarchically calculate the estimation losses of sub-regions, which alleviates the training bias. Extensive experiments on four challenging datasets (ShanghaiTech Part A, UCF_CC_50, UCF-QNRF, and WorldExpo'10) demonstrate the superiority of the proposed approach.

****

FocalMix: Semi-Supervised Learning for 3D Medical Image Detection

Dong Wang, Yuan Zhang, Kexin Zhang, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3951-3960

Applying artificial intelligence techniques in medical imaging is one of the most promising areas in medicine. However, most of the recent success in this area

highly relies on large amounts of carefully annotated data, whereas annotating m
edical images is a costly process. In this paper, we propose a novel method, cal
led FocalMix, which, to the best of our knowledge, is the first to leverage rece
nt advances in semi-supervised learning (SSL) for 3D medical image detection. We
 conducted extensive experiments on two widely used datasets for lung nodule det
ection, LUNA16 and NLST. Results show that our proposed SSL methods can achieve
 a substantial improvement of up to 17.3% over state-of-the-art supervised learni
ng approaches with 400 unlabeled CT scans.
****

Bi-Directional Relationship Inferring Network for Referring Image Segmentation
Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, Huchuan Lu; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp
. 4424-4433
Most existing methods do not explicitly formulate the mutual guidance between vi
sion and language. In this work, we propose a bi-directional relationship inferr
ing network (BRINet) to model the dependencies of cross-modal information. In de
tail, the vision-guided linguistic attention is used to learn the adaptive lingu
istic context corresponding to each visual region. Combining with the language-g
uided visual attention, a bi-directional cross-modal attention module (BCAM) is
built to learn the relationship between multi-modal features. Thus, the ultimate
 semantic context of the target object and referring expression can be represent
ed accurately and consistently. Moreover, a gated bi-directional fusion module (
GBFM) is designed to integrate the multi-level features where a gate function is
 used to guide the bi-directional flow of multi-level information. Extensive exp
eriments on four benchmark datasets demonstrate that the proposed method outperf
orms other state-of-the-art methods under different evaluation metrics.
****

FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation
Matias Tassano, Julie Delon, Thomas Veit; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1354-1363
In this paper, we propose a state-of-the-art video denoising algorithm based on
a convolutional neural network architecture. Until recently, video denoising wit
h neural networks had been a largely under explored domain, and existing methods
 could not compete with the performance of the best patch-based methods. The app
roach we introduce in this paper, called FastDVDnet, shows similar or better per
formance than other state-of-the-art competitors with significantly lower comput
ing times. In contrast to other existing neural network denoisers, our algorithm
 exhibits several desirable properties such as fast runtimes, and the ability to
 handle a wide range of noise levels with a single network model. The characteri
stics of its architecture make it possible to avoid using a costly motion compen
sation stage while achieving excellent performance. The combination between its
denoising performance and lower computational load makes this algorithm attracti
ve for practical denoising applications. We compare our method with different st
ate-of-art algorithms, both visually and with respect to objective quality metri
cs.
****

Composed Query Image Retrieval Using Locally Bounded Features
Mehrdad Hosseinzadeh, Yang Wang; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2020, pp. 3596-3605
Composed query image retrieval is a new problem where the query consists of an i
mage together with a requested modification expressed via a textual sentence. Th
e goal is then to retrieve the images that are generally similar to the query im
age, but differ according to the requested modification. Previous methods usuall
y consider the image as a whole. In this paper, we propose a novel method that r
epresents the image using a set of local areas in the image. The relationship be
tween each word in the modification text and each area in the image is then expl
icitly established, allowing the model to accurately correlate the modification
text to parts of the image. We conduct extensive experiments on three benchmark
datasets. The results show that our method outperforms other state-of-the-art ap
proaches by a considerable margin.

****
Variational-EM-Based Deep Learning for Noise-Blind Image Deblurring
Yuesong Nan, Yuhui Quan, Hui Ji; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2020, pp. 3626-3635
Non-blind deblurring is an important problem encountered in many image restorati on tasks. The focus of non-blind deblurring is on how to suppress noise magnific ation during deblurring. In practice, it often happens that the noise level of i nput image is unknown and varies among different images. This paper aims at deve loping a deep learning framework for deblurring images with unknown noise level. Based on the framework of variational expectation maximization (EM), an iterati ve noise-blind deblurring scheme is proposed which integrates the estimation of noise level and the quantification of image prior uncertainty. Then, the propose d scheme is unrolled to a neural network (NN) where image prior is modeled by NN with uncertainty quantification. Extensive experiments showed that the proposed method not only outperformed existing noise-blind deblurring methods by a large margin, but also outperformed those state-of-the-art image deblurring methods d esigned/trained with known noise level.
****
Central Similarity Quantization for Efficient Image and Video Retrieval
Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, Jia shi Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3083-3092
Existing data-dependent hashing methods usually learn hash functions from pairwi se or triplet data relationships, which only capture the data similarity locally , and often suffer from low learning efficiency and low collision rate. In this work, we propose a new global similarity metric, termed as central similarity, w ith which the hash codes of similar data pairs are encouraged to approach a comm on center and those for dissimilar pairs to converge to different centers, to im prove hash learning efficiency and retrieval accuracy. We principally formulate the computation of the proposed central similarity metric by introducing a new c oncept, i.e., hash center that refers to a set of data points scattered in the H amming space with a sufficient mutual distance between each other. We then provi de an efficient method to construct well separated hash centers by leveraging th e Hadamard matrix and Bernoulli distributions. Finally, we propose the Central S imilarity Quantization (CSQ) that optimizes the central similarity between data points w.r.t. their hash centers instead of optimizing the local similarity. CSQ is generic and applicable to both image and video hashing scenarios. Extensive experiments on large-scale image and video retrieval tasks demonstrate that CSQ can generate cohesive hash codes for similar data pairs and dispersed hash codes for dissimilar pairs, achieving a noticeable boost in retrieval performance, i. e. 3%-20% in mAP over the previous state-of-the-arts.
****
Taking a Deeper Look at Co-Salient Object Detection
Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, Ming-Ming C heng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2020, pp. 2919-2929
Co-salient object detection (CoSOD) is a newly emerging and rapidly growing bran ch of salient object detection (SOD), which aims to detect the co-occurring sali ent objects in multiple images. However, existing CoSOD datasets often have a se rious data bias, which assumes that each group of images contains salient object s of similar visual appearances. This bias results in the ideal settings and the effectiveness of the models, trained on existing datasets, may be impaired in r eal-life situations, where the similarity is usually semantic or conceptual. To tackle this issue, we first collect a new high-quality dataset, named CoSOD3k, w hich contains 3,316 images divided into 160 groups with multiple level annotatio ns, i.e., category, bounding box, object, and instance levels. CoSOD3k makes a s ignificant leap in terms of diversity, difficulty and scalability, benefiting re lated vision tasks. Besides, we comprehensively summarize 34 cutting-edge algori thms, benchmarking 19 of them over four existing CoSOD datasets (MSRC, iCoSeg, I mage Pair and CoSal2015) and our CoSOD3k with a total of 61K images (largest sc

ale), and reporting group-level performance analysis. Finally, we discuss the ch allenge and future work of CoSOD. Our study would give a strong boost to growth in the CoSOD community. Benchmark toolbox and results are available on our proje ct page.
****
Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics
Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, Siwei Lyu; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 320 7-3216
AI-synthesized face-swapping videos, commonly known as DeepFakes, is an emerging problem threatening the trustworthiness of online information. The need to deve lop and evaluate DeepFake detection algorithms calls for datasets of DeepFake vi deos. However, current DeepFake datasets suffer from low visual quality and do n ot resemble DeepFake videos circulated on the Internet. We present a new large-s cale challenging DeepFake video dataset, Celeb-DF, which contains 5,639 high-qua lity DeepFake videos of celebrities generated using improved synthesis process. We conduct a comprehensive evaluation of DeepFake detection methods and datasets to demonstrate the escalated level of challenges posed by Celeb-DF.
****
TEA: Temporal Excitation and Aggregation for Action Recognition
Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, Limin Wang; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 909-918
Temporal modeling is key for action recognition in videos. It normally considers both short-range motions and long-range aggregations. In this paper, we propose a Temporal Excitation and Aggregation (TEA) block, including a motion excitatio n (ME) module and a multiple temporal aggregation (MTA) module, specifically des igned to capture both short- and long-range temporal evolution. In particular, f or short-range motion modeling, the ME module calculates the feature-level tempo ral differences from spatiotemporal features. It then utilizes the differences t o excite the motion-sensitive channels of the features. The long-range temporal aggregations in previous works are typically achieved by stacking a large number of local temporal convolutions. Each convolution processes a local temporal win dow at a time. In contrast, the MTA module proposes to deform the local convolut ion to a group of sub-convolutions, forming a hierarchical residual architecture . Without introducing additional parameters, the features will be processed with a series of sub-convolutions, and each frame could complete multiple temporal a ggregations with neighborhoods. The final equivalent receptive field of temporal dimension is accordingly enlarged, which is capable of modeling the long-range temporal relationship over distant frames. The two components of the TEA block a re complementary in temporal modeling. Finally, our approach achieves impressive results at low FLOPs on several action recognition benchmarks, such as Kinetics , Something-Something, HMDB51, and UCF101, which confirms its effectiveness and efficiency.
****
Unsupervised Person Re-Identification via Softened Similarity Learning
Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, Qi Tian; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3 390-3399
Person re-identification (re-ID) is an important topic in computer vision. This paper studies the unsupervised setting of re-ID, which does not require any labe led information and thus is freely deployed to new scenarios. There are very few studies under this setting, and one of the best approach till now used iterativ e clustering and classification, so that unlabeled images are clustered into pse udo classes for a classifier to get trained, and the updated features are used f or clustering and so on. This approach suffers two problems, namely, the difficu lty of determining the number of clusters, and the hard quantization loss in clu stering. In this paper, we follow the iterative training mechanism but discard c lustering, since it incurs loss from hard quantization, yet its only product, im age-level similarity, can be easily replaced by pairwise computation and a softe

ned classification task. With these improvements, our approach becomes more eleg ant and is more robust to hyper-parameter changes. Experiments on two image-base d and video-based datasets demonstrate state-of-the-art performance under the un supervised re-ID setting.
****

Frequency Domain Compact 3D Convolutional Neural Networks
Hanting Chen, Yunhe Wang, Han Shu, Yehui Tang, Chunjing Xu, Boxin Shi, Cha o Xu, Qi Tian, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2020, pp. 1641-1650
This paper studies the compression and acceleration of 3-dimensional convolution al neural networks (3D CNNs). To reduce the memory cost and computational comple xity of deep neural networks, a number of algorithms have been explored by disco vering redundant parameters in pre-trained networks. However, most of existing m ethods are designed for processing neural networks consisting of 2-dimensional c onvolution filters (i.e. image classification and detection) and cannot be strai ghtforwardly applied for 3-dimensional filters (i.e. time series data). In this paper, we develop a novel approach for eliminating redundancy in the time dimens ionality of 3D convolution filters by converting them into the frequency domain through a series of learned optimal transforms with extremely fewer parameters. Moreover, these transforms are forced to be orthogonal, and the calculation of f eature maps can be accomplished in the frequency domain to achieve considerable speed-up rates. Experimental results on benchmark 3D CNN models and datasets dem onstrate that the proposed Frequency Domain Compact 3D CNNs (FDC3D) can achieve the state-of-the-art performance, e.g. a 2x speed-up ratio on the 3D-ResNet-18 w ithout obviously affecting its accuracy.
****

Revisiting Saliency Metrics: Farthest-Neighbor Area Under Curve
Sen Jia, Neil D. B. Bruce; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2020, pp. 2667-2676
In this paper, we propose a new metric to address the long-standing problem of c enter bias in saliency evaluation. We first show that distribution-based metrics cannot measure saliency performance across datasets due to ambiguity in the cho ice of standard deviation, especially for Convolutional Neural Networks. Therefo re, our proposed metric is AUC-based because ROC curves are relatively robust to the standard deviation problem. However, this requires sufficient unique values in the saliency prediction to compute AUC scores. Secondly, we propose a global smoothing function for the problem of few value degrees in predicted saliency o utput. Compared with random noise, our smoothing function can create unique valu es without losing the existing relative saliency relationship. Finally, we show our proposed AUC-based metric can generate a more directional negative set for e valuation, denoted as Farthest-Neighbor AUC (FN-AUC). Our experiments show FN-AU C can measure spatial biases, central and peripheral, more effectively than S-AU C without penalizing the fixation locations.
****

Structured Compression by Weight Encryption for Unstructured Pruning and Quantiz ation
Se Jung Kwon, Dongsoo Lee, Byeongwook Kim, Parichay Kapoor, Baeseong Park, Gu-Yeon Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2020, pp. 1909-1918
Model compression techniques, such as pruning and quantization, are becoming inc reasingly important to reduce the memory footprints and the amount of computatio ns. Despite model size reduction, achieving performance enhancement on devices i s, however, still challenging mainly due to the irregular representations of spa rse matrix formats. This paper proposes a new weight representation scheme for S parse Quantized Neural Networks, specifically achieved by fine-grained and unstr uctured pruning method. The representation is encrypted in a structured regular format, which can be efficiently decoded through XOR-gate network during inferen ce in a parallel manner. We demonstrate various deep learning models that can be compressed and represented by our proposed format with fixed and high compressi on ratio. For example, for fully-connected layers of AlexNet on ImageNet dataset

, we can represent the sparse weights by only 0.28 bits/weight for 1-bit quantiz
ation and 91% pruning rate with a fixed decoding rate and full memory bandwidth
usage. Decoding through XOR-gate network can be performed without any model accu
racy degradation with additional patch data associated with small overhead.
****

ARCH: Animatable Reconstruction of Clothed Humans
Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, Tony Tung; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,
pp. 3093-3102
In this paper, we propose ARCH (Animatable Reconstruction of Clothed Humans), a
novel end-to-end framework for accurate reconstruction of animation-ready 3D clo
thed humans from a monocular image. Existing approaches to digitize 3D humans st
ruggle to handle pose variations and recover details. Also, they do not produce
models that are animation ready. In contrast, ARCH is a learnedpose-awaremodelth
atproducesdetailed3Drigged full-body human avatars from a single unconstrained R
GB image. A Semantic Space and a Semantic Deformation Field are created using a
parametric 3D body estimator. They allow the transformation of 2D/3D clothed hum
ans into a canonical space, reducing ambiguities in geometry caused by pose vari
ations and occlusions in training data. Detailed surface geometry and appearance
are learned using an implicit function representation with spatial local featur
es. Furthermore, we propose additional per-pixel supervisiononthe3Dreconstructio
nusingopacity-awaredifferentiablerendering. OurexperimentsindicatethatARCH incre
ases the fidelity of the reconstructed humans. We obtain more than 50% lower rec
onstruction errors for standard metrics compared to state-of-the-art methods on
public datasets. We also show numerous qualitative examples of animated, high-qu
ality reconstructed avatars unseen in the literature so far.
****

Deep Implicit Volume Compression
Danhang Tang, Saurabh Singh, Philip A. Chou, Christian Hane, Mingsong Dou,
Sean Fanello, Jonathan Taylor, Philip Davidson, Onur G. Guleryuz, Yinda Zhan
g, Shahram Izadi, Andrea Tagliasacchi, Sofien Bouaziz, Cem Keskin; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2020, pp. 1293-1303
We describe a novel approach for compressing truncated signed distance fields (T
SDF) stored in 3D voxel grids, and their corresponding textures. To compress the
TSDF, our method relies on a block-based neural network architecture trained en
d-to-end, achieving state-of-the-art rate-distortion trade-off. To prevent topol
ogical errors, we losslessly com- press the signs of the TSDF, which also upper
bounds the reconstruction error by the voxel size. To compress the corresponding
texture, we designed a fast block-based UV parameterization, generating coheren
t texture maps that can be effectively compressed using existing video compressi
on algorithms. We demonstrate the performance of our algo- rithms on two 4D perf
ormance capture datasets, reducing bitrate by 66% for the same distortion, or al
ternatively re- ducing the distortion by 50% for the same bitrate, compared to t
he state-of-the-art.
****

Multi-Domain Learning for Accurate and Few-Shot Color Constancy
Jin Xiao, Shuhang Gu, Lei Zhang; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2020, pp. 3258-3267
Color constancy is an important process in camera pipeline to remove the color b
ias of captured image caused by scene illumination. Recently, significant improv
ements in color constancy accuracy have been achieved by using deep neural netwo
rks (DNNs). However, existing DNNbased color constancy methods learn distinct ma
ppings for different cameras, which require a costly data acquisition process fo
r each camera device. In this paper, we start a pioneer work to introduce multi-
domain learning to color constancy area. For different camera devices, we train
a branch of networks which share the same feature extractor and illuminant estim
ator, and only employ a camera-specific channel re-weighting module to adapt to
the camera-specific characteristics. Such a multi-domain learning strategy enabl
es us to take benefit from crossdevice training data. The proposed multi-domain

learning color constancy method achieved state-of-the-art performance on three c
ommonly used benchmark datasets. Furthermore, we also validate the proposed meth
od in a fewshot color constancy setting. Given a new unseen device with limited
number of training samples, our method is capable of delivering accurate color c
onstancy by merely learning the camera-specific parameters from the few-shot dat
aset. Our project page is publicly available at https://github.com/msxiaojin/MDL
CC.
****

Computing the Testing Error Without a Testing Set
Ciprian A. Corneanu, Sergio Escalera, Aleix M. Martinez; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2
677-2685
Deep Neural Networks (DNNs) have revolutionized computer vision. We now have DNN
s that achieve top (accuracy) results in many problems, including object recogni
tion, facial expression analysis, and semantic segmentation, to name but a few.
The design of the DNNs that achieve top results is, however, non-trivial and mos
tly done by trail-and-error. That is, typically, researchers will derive many DN
N architectures (i.e., topologies) and then test them on multiple datasets. Howe
ver, there are no guarantees that the selected DNN will perform well in the real
world. One can use a testing set to estimate the performance gap between the tr
aining and testing sets, but avoiding overfitting-to-the-testing-data is of conc
ern. Using a sequestered testing data may address this problem, but this require
s a constant update of the dataset, a very expensive venture. Here, we derive an
algorithm to estimate the performance gap between training and testing without
the need of a testing dataset. Specifically, we derive a set of persistent topol
ogy measures that identify when a DNN is learning to generalize to unseen sample
s. We provide extensive experimental validation on multiple networks and dataset
s to demonstrate the feasibility of the proposed approach.
****

Conditional Channel Gated Networks for Task-Aware Continual Learning
Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucc
hiara, Babak Ehteshami Bejnordi; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2020, pp. 3931-3940
Convolutional Neural Networks experience catastrophic forgetting when optimized
on a sequence of learning problems: as they meet the objective of the current tr
aining examples, their performance on previous tasks drops drastically. In this
work, we introduce a novel framework to tackle this problem with conditional com
putation. We equip each convolutional layer with task-specific gating modules, s
electing which filters to apply on the given input. This way, we achieve two app
ealing properties. Firstly, the execution patterns of the gates allow to identif
y and protect important filters, ensuring no loss in the performance of the mode
l for previously learned tasks. Secondly, by using a sparsity objective, we can
promote the selection of a limited set of kernels, allowing to retain sufficient
model capacity to digest new tasks. Existing solutions require, at test time, a
wareness of the task to which each example belongs to. This knowledge, however,
may not be available in many practical scenarios. Therefore, we additionally int
roduce a task classifier that predicts the task label of each example, to deal w
ith settings in which a task oracle is not available. We validate our proposal o
n four continual learning datasets. Results show that our model consistently out
performs existing methods both in the presence and the absence of a task oracle.
Notably, on Split SVHN and Imagenet-50 datasets, our model yields up to 23.98%
and 17.42% improvement in accuracy w.r.t. competing methods.
****

Polishing Decision-Based Adversarial Noise With a Customized Sampling
Yucheng Shi, Yahong Han, Qi Tian; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2020, pp. 1030-1038
As an effective black-box adversarial attack, decision-based methods polish adve
rsarial noise by querying the target model. Among them, boundary attack is widel
y applied due to its powerful noise compression capability, especially when comb
ined with transfer-based methods. Boundary attack splits the noise compression i

nto several independent sampling processes, repeating each query with a constant sampling setting. In this paper, we demonstrate the advantage of using current noise and historical queries to customize the variance and mean of sampling in boundary attack to polish adversarial noise. We further reveal the relationship between the initial noise and the compressed noise in boundary attack. We propose Customized Adversarial Boundary (CAB) attack that uses the current noise to model the sensitivity of each pixel and polish adversarial noise of each image with a customized sampling setting. On the one hand, CAB uses current noise as a prior belief to customize the multivariate normal distribution. On the other hand, CAB keeps the new samplings away from historical failed queries to avoid similar mistakes. Experimental results measured on several image classification datasets emphasizes the validity of our method.

****

G2L-Net: Global to Local Network for Real-Time 6D Pose Estimation With Embedding Vector Features

Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Ales Leonardis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4233-4242

In this paper, we propose a novel real-time 6D object pose estimation framework, named G2L-Net. Our network operates on point clouds from RGB-D detection in a divide-and-conquer fashion. Specifically, our network consists of three steps. First, we extract the coarse object point cloud from the RGB-D image by 2D detection. Second, we feed the coarse object point cloud to a translation localization network to perform 3D segmentation and object translation prediction. Third, via the predicted segmentation and translation, we transfer the fine object point cloud into a local canonical coordinate, in which we train a rotation localization network to estimate initial object rotation. In the third step, we define point-wise embedding vector features to capture viewpoint-aware information. To calculate more accurate rotation, we adopt a rotation residual estimator to estimate the residual between initial rotation and ground truth, which can boost initial pose estimation performance. Our proposed G2L-Net is real-time despite the fact multiple steps are stacked via the proposed coarse-to-fine framework. Extensive experiments on two benchmark datasets show that G2L-Net achieves state-of-the-art performance in terms of both accuracy and speed.

****

Pattern-Structure Diffusion for Multi-Task Learning

Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4514-4523

Inspired by the observation that pattern structures high-frequently recur within intra-task also across tasks, we propose a pattern-structure diffusion (PSD) framework to mine and propagate task-specific and task-across pattern structures in the task-level space for joint depth estimation, segmentation and surface normal prediction. To represent local pattern structures, we model them as small-scale graphlets, and propagate them in two different ways, i.e., intra-task and inter-task PSD. For the former, to overcome the limit of the locality of pattern structures, we use the high-order recursive aggregation on neighbors to multiplicatively increase the spread scope, so that long-distance patterns are propagated in the intra-task space. In the inter-task PSD, we mutually transfer the counterpart structures corresponding to the same spatial position into the task itself based on the matching degree of paired pattern structures therein. Finally, the intra-task and inter-task pattern structures are jointly diffused among the task-level patterns, and encapsulated into an end-to-end PSD network to boost the performance of multi-task learning. Extensive experiments on two widely-used benchmarks demonstrate that our proposed PSD is more effective and also achieves the state-of-the-art or competitive results.

****

On the Acceleration of Deep Learning Model Parallelism With Staleness

An Xu, Zhouyuan Huo, Heng Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2088-2097

Training the deep convolutional neural network for computer vision problems is s
low and inefficient, especially when it is large and distributed across multiple
devices. The inefficiency is caused by the backpropagation algorithm's forward
locking, backward locking, and update locking problems. Existing solutions for a
cceleration either can only handle one locking problem or lead to severe accurac
y loss or memory inefficiency. Moreover, none of them consider the straggler pro
blem among devices. In this paper, we propose Layer-wise Staleness and a novel e
fficient training algorithm, Diversely Stale Parameters (DSP), to address these
challenges. We also analyze the convergence of DSP with two popular gradient-bas
ed methods and prove that both of them are guaranteed to converge to critical po
ints for non-convex problems. Finally, extensive experimental results on trainin
g deep learning models demonstrate that our proposed DSP algorithm can achieve s
ignificant training speedup with stronger robustness than compared methods.
****

Self-Supervised Scene De-Occlusion
Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, Chen Change Loy;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2020, pp. 3784-3792
Natural scene understanding is a challenging task, particularly when encounterin
g images of multiple objects that are partially occluded. This obstacle is given
rise by varying object ordering and positioning. Existing scene understanding p
aradigms are able to parse only the visible parts, resulting in incomplete and u
nstructured scene interpretation. In this paper, we investigate the problem of s
cene de-occlusion, which aims to recover the underlying occlusion ordering and c
omplete the invisible parts of occluded objects. We make the first attempt to ad
dress the problem through a novel and unified framework that recovers hidden sce
ne structures without ordering and amodal annotations as supervisions. This is a
chieved via Partial Completion Network (PCNet)-mask (M) and -content (C), that l
earn to recover fractions of object masks and contents, respectively, in a self-
supervised manner. Based on PCNet-M and PCNet-C, we devise a novel inference sch
eme to accomplish scene de-occlusion, via progressive ordering recovery, amodal
completion and content completion. Extensive experiments on real-world scenes de
monstrate the superior performance of our approach to other alternatives. Remark
ably, our approach that is trained in a self-supervised manner achieves comparab
le results to fully-supervised methods. The proposed scene de-occlusion framewor
k benefits many applications, including high-quality and controllable image mani
pulation and scene recomposition (see Fig. 1), as well as the conversion of exis
ting modal mask annotations to amodal mask annotations.
****

DeepFLASH: An Efficient Network for Learning-Based Medical Image Registration
Jian Wang, Miaomiao Zhang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2020, pp. 4444-4452
This paper presents DeepFLASH, a novel network with efficient training and infer
ence for learning-based medical image registration. In contrast to existing appr
oaches that learn spatial transformations from training data in the high dimensi
onal imaging space, we develop a new registration network entirely in a low dime
nsional bandlimited space. This dramatically reduces the computational cost and
memory footprint of an expensive training and inference. To achieve this goal, w
e first introduce complex-valued operations and representations of neural archit
ectures that provide key components for learning-based registration models. We t
hen construct an explicit loss function of transformation fields fully character
ized in a bandlimited space with much fewer parameterizations. Experimental resu
lts show that our method is significantly faster than the state-of-the-art deep
learning based image registration methods, while producing equally accurate alig
nment. We demonstrate our algorithm in two different applications of image regis
tration: 2D synthetic data and 3D real brain magnetic resonance (MR) images.
****

Structure Boundary Preserving Segmentation for Medical Image With Ambiguous Boun
dary
Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, Yong Man Ro; Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4817-4826
In this paper, we propose a novel image segmentation method to tackle two critical problems of medical image, which are (i) ambiguity of structure boundary in the medical image domain and (ii) uncertainty of the segmented region without specialized domain knowledge. To solve those two problems in automatic medical segmentation, we propose a novel structure boundary preserving segmentation framework. To this end, the boundary key point selection algorithm is proposed. In the proposed algorithm, the key points on the structural boundary of the target object are estimated. Then, a boundary preserving block (BPB) with the boundary key point map is applied for predicting the structure boundary of the target object. Further, for embedding experts' knowledge in the fully automatic segmentation, we propose a novel shape boundary-aware evaluator (SBE) with the ground-truth structure information indicated by experts. The proposed SBE could give feedback to the segmentation network based on the structure boundary key point. The proposed method is general and flexible enough to be built on top of any deep learning-based segmentation network. We demonstrate that the proposed method could surpass the state-of-the-art segmentation network and improve the accuracy of three different segmentation network models on different types of medical image datasets.

****

Residual Feature Aggregation Network for Image Super-Resolution
Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, Gangshan Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2359-2368
Recently, very deep convolutional neural networks (CNNs) have shown great power in single image super-resolution (SISR) and achieved significant improvements against traditional methods. Among these CNN-based methods, the residual connections play a critical role in boosting the network performance. As the network depth grows, the residual features gradually focused on different aspects of the input image, which is very useful for reconstructing the spatial details. However, existing methods neglect to fully utilize the hierarchical features on the residual branches. To address this issue, we propose a novel residual feature aggregation (RFA) framework for more efficient feature extraction. The RFA framework groups several residual modules together and directly forwards the features on each local residual branch by adding skip connections. Therefore, the RFA framework is capable of aggregating these informative residual features to produce more representative features. To maximize the power of the RFA framework, we further propose an enhanced spatial attention (ESA) block to make the residual features to be more focused on critical spatial contents. The ESA block is designed to be lightweight and efficient. Our final RFANet is constructed by applying the proposed RFA framework with the ESA blocks. Comprehensive experiments demonstrate the necessity of our RFA framework and the superiority of our RFANet over state-of-the-art SISR methods.

****

Context-Aware and Scale-Insensitive Temporal Repetition Counting
Huaidong Zhang, Xuemiao Xu, Guoqiang Han, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 670-678
Temporal repetition counting aims to estimate the number of cycles of a given repetitive action. Existing deep learning methods assume repetitive actions are performed in a fixed time-scale, which is invalid for the complex repetitive actions in real life. In this paper, we tailor a context-aware and scale-insensitive framework, to tackle the challenges in repetition counting caused by the unknown and diverse cycle-lengths. Our approach combines two key insights: (1) Cycle lengths from different actions are unpredictable that require large-scale searching, but, once a coarse cycle length is determined, the variety between repetitions can be overcome by regression. (2) Determining the cycle length cannot only rely on a short fragment of video but a contextual understanding. The first point is implemented by a coarse-to-fine cycle refinement method. It avoids the heavy

computation of exhaustively searching all the cycle lengths in the video, and, i nstead, it propagates the coarse prediction for further refinement in a hierarch ical manner. We secondly propose a bidirectional cycle length estimation method for a context-aware prediction. It is a regression network that takes two consec utive coarse cycles as input, and predicts the locations of the previous and nex t repetitive cycles. To benefit the training and evaluation of temporal repetiti on counting area, we construct a new and largest benchmark, which contains 526 v ideos with diverse repetitive actions. Extensive experiments show that the propo sed network trained on a single dataset outperforms state-of-the-art methods on several benchmarks, indicating that the proposed framework is general enough to capture repetition patterns across domains. Code and data are available in https ://github.com/Xiaodomgdomg/Deep-Temporal-Repetition-Counting.
****

DEPARA: Deep Attribution Graph for Deep Knowledge Transferability
Jie Song, Yixin Chen, Jingwen Ye, Xinchao Wang, Chengchao Shen, Feng Mao, Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2020, pp. 3922-3930
Exploring the intrinsic interconnections between the knowledge encoded in PRe-tr ained Deep Neural Networks (PR-DNNs) of heterogeneous tasks sheds light on their mutual transferability, and consequently enables knowledge transfer from one ta sk to another so as to reduce the training effort of the latter. In this paper, we propose the DEeP Attribution gRAph (DEPARA) to investigate the transferabilit y of knowledge learned from PR-DNNs. In DEPARA, nodes correspond to the inputs a nd are represented by their vectorized attribution maps with regards to the outp uts of the PR-DNN. Edges denote the relatedness between inputs and are measured by the similarity of their features extracted from the PR-DNN. The knowledge tra nsferability of two PR-DNNs is measured by the similarity of their corresponding DEPARAs. We apply DEPARA to two important yet under-studied problems in transfe r learning: pre-trained model selection and layer selection. Extensive experimen ts are conducted to demonstrate the effectiveness and superiority of the propose d method in solving both these problems. Code, data and models reproducing the r esults in this paper are available at https://github.com/zju-vipa/DEPARA.
****

Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognit ion
Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, Wanli Ouyang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2020, pp. 143-152
Spatial-temporal graphs have been widely used by skeleton-based action recogniti on algorithms to model human action dynamics. To capture robust movement pattern s from these graphs, long-range and multi-scale context aggregation and spatial-temporal dependency modeling are critical aspects of a powerful feature extracto r. However, existing methods have limitations in achieving (1) unbiased long-ran ge joint relationship modeling under multi-scale operators and (2) unobstructed cross-spacetime information flow for capturing complex spatial-temporal dependen cies. In this work, we present (1) a simple method to disentangle multi-scale gr aph convolutions and (2) a unified spatial-temporal graph convolutional operator named G3D. The proposed multi-scale aggregation scheme disentangles the importa nce of nodes in different neighborhoods for effective long-range modeling. The p roposed G3D module leverages dense cross-spacetime edges as skip connections for direct information propagation across the spatial-temporal graph. By coupling t hese proposals, we develop a powerful feature extractor named MS-G3D based on wh ich our model outperforms previous state-of-the-art methods on three large-scale datasets: NTU RGB+D 60, NTU RGB+D 120, and Kinetics Skeleton 400.
****

Category-Level Articulated Object Pose Estimation
Xiaolong Li, He Wang, Li Yi, Leonidas J. Guibas, A. Lynn Abbott, Shuran Son g; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2020, pp. 3706-3715
This paper addresses the task of category-level pose estimation for articulated

objects from a single depth image. We present a novel category-level approach th at correctly accommodates object instances previously unseen during training. We introduce Articulation-aware Normalized Coordinate Space Hierarchy (ANCSH) - a canonical representation for different articulated objects in a given category. As the key to achieve intra-category generalization, the representation construc ts a canonical object space as well as a set of canonical part spaces. The canon ical object space normalizes the object orientation, scales and articulations (e .g. joint parameters and states) while each canonical part space further normali zes its part pose and scale. We develop a deep network based on PointNet++ that predicts ANCSH from a single depth point cloud, including part segmentation, nor malized coordinates, and joint parameters in the canonical object space. By leve raging the canonicalized joints, we demonstrate: 1) improved performance in part pose and scale estimations using the induced kinematic constraints from joints; 2) high accuracy for joint parameter estimation in camera space.

****

ImVoteNet: Boosting 3D Object Detection in Point Clouds With Image Votes

Charles R. Qi, Xinlei Chen, Or Litany, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp . 4404-4413

3D object detection has seen quick progress thanks to advances in deep learning on point clouds. A few recent works have even shown state-of-the-art performance with just point clouds input (e.g. VoteNet). However, point cloud data have inh erent limitations. They are sparse, lack color information and often suffer from sensor noise. Images, on the other hand, have high resolution and rich texture. Thus they can complement the 3D geometry provided by point clouds. Yet how to e ffectively use image information to assist point cloud based detection is still an open question. In this work, we build on top of VoteNet and propose a 3D dete ction architecture called ImVoteNet specialized for RGB-D scenes. ImVoteNet is b ased on fusing 2D votes in images and 3D votes in point clouds. Compared to prio r work on multi-modal detection, we explicitly extract both geometric and semant ic features from the 2D images. We leverage camera parameters to lift these feat ures to 3D. To improve the synergy of 2D-3D feature fusion, we also propose a mu lti-tower training scheme. We validate our model on the challenging SUN RGB-D da taset, advancing state-of-the-art results by 5.7 mAP. We also provide rich ablat ion studies to analyze the contribution of each design choice.

****

Achieving Robustness in the Wild via Adversarial Mixing With Disentangled Repres entations

Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotha m, Timothy Mann, Pushmeet Kohli; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2020, pp. 1211-1220

Recent research has made the surprising finding that state-of-the-art deep learn ing models sometimes fail to generalize to small variations of the input. Advers arial training has been shown to be an effective approach to overcome this probl em. However, its application has been limited to enforcing invariance to analyti cally defined transformations like lp-norm bounded perturbations. Such perturbat ions do not necessarily cover plausible real-world variations that preserve the semantics of the input (such as a change in lighting conditions). In this paper, we propose a novel approach to express and formalize robustness to these kinds of real-world transformations of the input. The two key ideas underlying our for mulation are (1) leveraging disentangled representations of the input to define different factors of variations, and (2) generating new input images by adversar ially composing the representations of different images. We use a StyleGAN model to demonstrate the efficacy of this framework. Specifically, we leverage the di sentangled latent representations computed by a StyleGAN model to generate pertu rbations of an image that are similar to real-world variations (like adding make -up, or changing the skin-tone of a person) and train models to be invariant to these perturbations. Extensive experiments show that our method improves general ization and reduces the effect of spurious correlations (reducing the error rate of a "smile" detector by 21% for example).

****
Deep Non-Line-of-Sight Reconstruction
Javier Grau Chopite, Matthias B. Hullin, Michael Wand, Julian Iseringhausen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2020, pp. 960-969
The recent years have seen a surge of interest in methods for imaging beyond the direct line of sight. The most prominent techniques rely on time-resolved optic al impulse responses, obtained by illuminating a diffuse wall with an ultrashort light pulse and observing multi-bounce indirect reflections with an ultrafast t ime-resolved imager. Reconstruction of geometry from such data, however, is a co mplex non-linear inverse problem that comes with substantial computational deman ds. In this paper, we employ convolutional feed-forward networks for solving the reconstruction problem efficiently while maintaining good reconstruction qualit y. Specifically, we devise a tailored autoencoder architecture, trained end-to-e nd, that maps transient images directly to a depth-map representation. Training is done using a recent, very efficient transient renderer for three-bounce indir ect light transport that enables the quick generation of large amounts of traini ng data for the network. We examine the performance of our method on a variety o f synthetic and experimental datasets and its dependency on the choice of traini ng data and augmentation strategies, as well as architectural features. We demon strate that our feed-forward network, even if trained solely on synthetic data, is able to obtain results competitive with previous, model-based optimization me thods, while being orders of magnitude faster.
****
Unsupervised Adaptation Learning for Hyperspectral Imagery Super-Resolution
Lei Zhang, Jiangtao Nie, Wei Wei, Yanning Zhang, Shengcai Liao, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2020, pp. 3073-3082
The key for fusion based hyperspectral image (HSI) super-resolution (SR) is to i nfer the posteriori of a latent HSI using appropriate image prior and likelihood that depends on degeneration. However, in practice the priors of high-dimension al HSIs can be extremely complicated and the degeneration is often unknown. Cons equently most existing approaches that assume a shallow hand-crafted image prior and a pre-defined degeneration, fail to well generalize in real applications. T o tackle this problem, we present an unsupervised adaptation learning (UAL) fram ework. Instead of directly modelling the complicated image prior, we propose to first implicitly learn a general image prior using deep networks and then adapt it to a specific HSI. Following this idea, we develop a two-stage SR network tha t leverages two consecutive modules: a fusion module and an adaptation module, t o recover the latent HSI in a coarse-to-fine scheme. The fusion module is pretra ined in a supervised manner on synthetic data to capture a spatial-spectral prio r that is general across most HSIs. To adapt the learned general prior to the sp ecific HSI under unknown degeneration, we introduce a simple degeneration networ k to assist learning both the adaptation module and the degeneration in an unsup ervised way. In this way, the resultant image-specific prior and the estimated d egeneration can benefit the inference of a more accurate posteriori, thereby inc reasing generalization capacity. To verify the efficacy of UAL, we extensively e valuate it on four benchmark datasets and report strong results that surpass exi sting approaches.
****
Joint Demosaicing and Denoising With Self Guidance
Lin Liu, Xu Jia, Jianzhuang Liu, Qi Tian; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2240-2249
Usually located at the very early stages of the computational photography pipeli ne, demosaicing and denoising play important parts in the modern camera image pr ocessing. Recently, some neural networks have shown the effectiveness in joint d emosaicing and denoising (JDD). Most of them first decompose a Bayer raw image i nto a four-channel RGGB image and then feed it into a neural network. This pract ice ignores the fact that the green channels are sampled at a double rate compar ed to the red and the blue channels. In this paper, we propose a self-guidance n

etwork (SGNet), where the green channels are initially estimated and then works as a guidance to recover all missing values in the input image. In addition, as regions of different frequencies suffer different levels of degradation in image restoration. We propose a density-map guidance to help the model deal with a wide range of frequencies. Our model outperforms state-of-the-art joint demosaicing and denoising methods on four public datasets, including two real and two synthetic data sets. Finally, we also verify that our method obtains best results in joint demosaicing , denoising and super-resolution.
****

SpSequenceNet: Semantic Segmentation Network on 4D Point Clouds
Hanyu Shi, Guosheng Lin, Hao Wang, Tzu-Yi Hung, Zhenhua Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4574-4583
Point clouds are useful in many applications like autonomous driving and robotics as they provide natural 3D information of the surrounding environments. While there are extensive research on 3D point clouds, scene understanding on 4D point clouds, a series of consecutive 3D point clouds frames, is an emerging topic and yet under-investigated. With 4D point clouds (3D point cloud videos), robotic systems could enhance their robustness by leveraging the temporal information from previous frames. However, the existing semantic segmentation methods on 4D point clouds suffer from low precision due to the spatial and temporal information loss in their network structures. In this paper, we propose SpSequenceNet to address this problem. The network is designed based on 3D sparse convolution. And we introduce two novel modules, a cross-frame global attention module and a cross-frame local interpolation module, to capture spatial and temporal information in 4D point clouds. We conduct extensive experiments on SemanticKITTI, and achieve the state-of-the-art result of 43.1% on mIoU, which is 1.5% higher than the previous best approach.
****

Shoestring: Graph-Based Semi-Supervised Classification With Severely Limited Labeled Data
Wanyu Lin, Zhaolin Gao, Baochun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4174-4182
Graph-based semi-supervised learning has been shown to be one of the most effective classification approaches, as it can exploit connectivity patterns between labeled and unlabeled samples to improve learning performance. However, we show that existing techniques perform poorly when labeled data are severely limited. To address the problem of semi-supervised learning in the presence of severely limited labeled samples, we propose a new framework, called Shoestring , that incorporates metric learning into the paradigm of graph-based semi-supervised learning. In particular, our base model consists of a graph embedding network, followed by a metric learning network that learns a semantic metric space to represent the semantic similarity between the sparsely labeled and large numbers of unlabeled samples. Then the classification can be performed by clustering the unlabeled samples according to the learned semantic space. We empirically demonstrate Shoestring's superiority over many baselines, including graph convolutional networks, label propagation and their recent label-efficient variations (IGCN and GLP). We show that our framework achieves state-of-the-art performance for node classification in the low-data regime. In addition, we demonstrate the effectiveness of our framework on image classification tasks in the few-shot learning regime, with significant gains on miniImageNet (2.57%~3.59%) and tieredImageNet (1.05%~2.70%).
****

Distilling Image Dehazing With Heterogeneous Task Imitation
Ming Hong, Yuan Xie, Cuihua Li, Yanyun Qu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3462-3471
State-of-the-art deep dehazing models are often difficult in training. Knowledge distillation paves a way to train a student network assisted by a teacher network. However, most knowledge distill methods are used for image classification and segmentation as well as object detection, and few investigate distilling image

restoration and use different task for knowledge transfer. In this paper, we propose a knowledge-distill dehazing network which distills image dehazing with the heterogeneous task imitation. In our network, the teacher is an off-the-shelf auto-encoder network and is used for image reconstruction. The dehazing network is trained assisted by the teacher network with the process-oriented learning mechanism. The student network imitates the task of image reconstruction in the teacher network. Moreover, we design a spatial-weighted channel-attention residual block for the student image dehazing network to adaptively learn the content-aware channel level attention and pay more attention to the features for dense hazy regions reconstruction. To evaluate the effectiveness of the proposed method, we compare our method with several state-of-the-art methods on two synthetic and real-world datasets, as well as real hazy images.

****

Photometric Stereo via Discrete Hypothesis-and-Test Search

Kenji Enomoto, Michael Waechter, Kiriakos N. Kutulakos, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2311-2319

In this paper, we consider the problem of estimating surface normals of a scene with spatially varying, general BRDFs observed by a static camera under varying, known, distant illumination. Unlike previous approaches that are mostly based on continuous local optimization, we cast the problem as a discrete hypothesis-and-test search problem over the discretized space of surface normals. While a naive search requires a significant amount of time, we show that the expensive computation block can be precomputed in a scene-independent manner, resulting in accelerated inference for new scenes. It allows us to perform a full search over the finely discretized space of surface normals to determine the globally optimal surface normal for each scene point. We show that our method can accurately estimate surface normals of scenes with spatially varying different reflectances in a reasonable amount of time.

****

Task Agnostic Robust Learning on Corrupt Outputs by Correlation-Guided Mixture Density Networks

Sungjoon Choi, Sanghoon Hong, Kyungjae Lee, Sungbin Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3872-3881

In this paper, we focus on weakly supervised learning with noisy training data for both classification and regression problems. We assume that the training outputs are collected from a mixture of a target and correlated noise distributions. Our proposed method simultaneously estimates the target distribution and the quality of each data which is defined as the correlation between the target and data generating distributions. The cornerstone of the proposed method is a Cholesky Block that enables modeling dependencies among mixture distributions in a differentiable manner where we maintain the distribution over the network weights. We first provide illustrative examples in both regression and classification tasks to show the effectiveness of the proposed method. Then, the proposed method is extensively evaluated in a number of experiments where we show that it constantly shows comparable or superior performances compared to existing baseline methods in the handling of noisy data.

****

Multi-Path Region Mining for Weakly Supervised 3D Semantic Segmentation on Point Clouds

Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, Lihua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4384-4393

Point clouds provide intrinsic geometric information and surface context for scene understanding. Existing methods for point cloud segmentation require a large amount of fully labeled data. Using advanced depth sensors, collection of large scale 3D dataset is no longer a cumbersome process. However, manually producing point-level label on the large scale dataset is time and labor-intensive. In this paper, we propose a weakly supervised approach to predict point-level results

using weak labels on 3D point clouds. We introduce our multi-path region mining module to generate pseudo point-level labels from a classification network trained with weak labels. It mines the localization cues for each class from various aspects of the network feature using different attention modules. Then, we use the point-level pseudo label to train a point cloud segmentation network in a fully supervised manner. To the best of our knowledge, this is the first method that uses cloud-level weak labels on raw 3D space to train a point cloud semantic segmentation network. In our setting, the 3D weak labels only indicate the classes that appeared in our input sample. We discuss both scene- and subcloud-level weakly labels on raw 3D point cloud data and perform in-depth experiments on them. On ScanNet dataset, our result trained with subcloud-level labels is compatible with some fully supervised methods.
****

Single-Step Adversarial Training With Dropout Scheduling
Vivek B.S., R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 950-959
Deep learning models have shown impressive performance across a spectrum of computer vision applications including medical diagnosis and autonomous driving. One of the major concerns that these models face is their susceptibility to adversarial attacks. Realizing the importance of this issue, more researchers are working towards developing robust models that are less affected by adversarial attacks. Adversarial training method shows promising results in this direction. In adversarial training regime, models are trained with mini-batches augmented with adversarial samples. Fast and simple methods (e.g., single-step gradient ascent) are used for generating adversarial samples, in order to reduce computational complexity. It is shown that models trained using single-step adversarial training method (adversarial samples are generated using non-iterative method) are pseudo robust. Further, this pseudo robustness of models is attributed to the gradient masking effect. However, existing works fail to explain when and why gradient masking effect occurs during single-step adversarial training. In this work, (i) we show that models trained using single-step adversarial training method learn to prevent the generation of single-step adversaries, and this is due to over-fitting of the model during the initial stages of training, and (ii) to mitigate this effect, we propose a single-step adversarial training method with dropout scheduling. Unlike models trained using existing single-step adversarial training methods, models trained using the proposed single-step adversarial training method are robust against both single-step and multi-step adversarial attacks, and the performance is on par with models trained using computationally expensive multi-step adversarial training methods, in white-box and black-box settings.
****

Online Depth Learning Against Forgetting in Monocular Videos
Zhenyu Zhang, Stephane Lathuiliere, Elisa Ricci, Nicu Sebe, Yan Yan, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4494-4503
Online depth learning is the problem of consistently adapting a depth estimation model to handle a continuously changing environment. This problem is challenging due to the network easily overfits on the current environment and forgets its past experiences. To address such problem, this paper presents a novel Learning to Prevent Forgetting (LPF) method for online mono-depth adaptation to new target domains in unsupervised manner. Instead of updating the universal parameters, LPF learns adapter modules to efficiently adjust the feature representation and distribution without losing the pre-learned knowledge in online condition. Specifically, to adapt temporal-continuous depth patterns in videos, we introduce a novel meta-learning approach to learn adapter modules by combining online adaptation process into the learning objective. To further avoid overfitting, we propose a novel temporal-consistent regularization to harmonize the gradient descent procedure at each online learning step. Extensive evaluations on real-world datasets demonstrate that the proposed method, with very limited parameters, significantly improves the estimation quality.
****

Neuromorphic Camera Guided High Dynamic Range Imaging
Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1730-1739
Reconstruction of high dynamic range image from a single low dynamic range image captured by a frame-based conventional camera, which suffers from over- or under-exposure, is an ill-posed problem. In contrast, recent neuromorphic cameras are able to record high dynamic range scenes in the form of an intensity map, with much lower spatial resolution, and without color. In this paper, we propose a neuromorphic camera guided high dynamic range imaging pipeline, and a network consisting of specially designed modules according to each step in the pipeline, which bridges the domain gaps on resolution, dynamic range, and color representation between two types of sensors and images. A hybrid camera system has been built to validate that the proposed method is able to reconstruct quantitatively and qualitatively high-quality high dynamic range images by successfully fusing the images and intensity maps for various real-world scenarios.
****

Reconstruct Locally, Localize Globally: A Model Free Method for Object Pose Estimation
Ming Cai, Ian Reid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3153-3163
Six degree-of-freedom pose estimation of a known object in a single image is a long-standing computer vision objective. It is classically posed as a correspondence problem between a known geometric model, such as a CAD model, and image locations. If a CAD model is not available, it is possible to use multi-view visual reconstruction methods to create a geometric model, and use this in the same manner. Instead, we propose a learning-based method whose input is a collection of images of a target object, and whose output is the pose of the object in a novel view. At inference time, our method maps from the RoI features of the input image to a dense collection of object-centric 3D coordinates, one per pixel. This dense 2D-3D mapping is then used to determine 6dof pose using standard PnP plus RANSAC. The model that maps 2D to object 3D coordinates is established at training time by automatically discovering and matching image landmarks that are consistent across multiple views. We show that this method eliminates the requirement for a 3D CAD model (needed by classical geometry-based methods and state-of-the-art learning-based methods alike) but still achieves performance on a par with the prior art.
****

Deep Learning for Handling Kernel/model Uncertainty in Image Deconvolution
Yuesong Nan, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2388-2397
Most existing non-blind image deconvolution methods assume that the given blurring kernel is error-free. In practice, blurring kernel often is estimated via some blind deblurring algorithm which is not exactly the truth. Also, the convolution model is only an approximation to practical blurring effect. It is known that non-blind deconvolution is susceptible to such a kernel/model error. Based on an error-in-variable (EIV) model of image blurring that takes kernel error into consideration, this paper presents a deep learning method for deconvolution, which unrolls a total-least-squares (TLS) estimator whose relating priors are learned by neural networks (NNs). The experiments showed that the proposed method is robust to kernel/model error. It noticeably outperformed existing solutions when deblurring images using noisy kernels, e.g. the ones estimated from existing blind motion deblurring methods.
****

VPLNet: Deep Single View Normal Estimation With Vanishing Points and Lines
Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, Jan-Michael Frahm; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 689-698
We present a novel single-view surface normal estimation method that combines traditional line and vanishing point analysis with a deep learning approach. Start

ing from a color image and a Manhattan line map, we use a deep neural network to regress on a dense normal map, and a dense Manhattan label map that identifies planar regions aligned with the Manhattan directions. We fuse the normal map and label map in a fully differentiable manner to produce a refined normal map as final output. To do so, we softly decompose the output into a Manhattan part and a non-Manhattan part. The Manhattan part is treated by discrete classification and vanishing points, while the non-Manhattan part is learned by direct supervision. Our method achieves state-of-the-art results on standard single-view normal estimation benchmarks. More importantly, we show that by using vanishing points and lines, our method has better generalization ability than existing works. In addition, we demonstrate how our surface normal network can improve the performance of depth estimation networks, both quantitatively and qualitatively, in particular, in 3D reconstructions of walls and other flat surfaces.
****

Intra- and Inter-Action Understanding via Temporal Action Parsing
Dian Shao, Yue Zhao, Bo Dai, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 730-739
Current methods for action recognition primarily rely on deep convolutional networks to derive feature embeddings of visual and motion features. While these methods have demonstrated remarkable performance on standard benchmarks, we are still in need of a better understanding as to how the videos, in particular their internal structures, relate to high-level semantics, which may lead to benefits in multiple aspects, e.g. interpretable predictions and even new methods that can take the recognition performances to a next level. Towards this goal, we construct TAPOS, a new dataset developed on sport videos with manual annotations of sub-actions, and conduct a study on temporal action parsing on top. Our study shows that a sport activity usually consists of multiple sub-actions and that the awareness of such temporal structures is beneficial to action recognition. We also investigate a number of temporal parsing methods, and thereon devise an improved method that is capable of mining sub-actions from training data without knowing the labels of them. On the constructed TAPOS, the proposed method is shown to reveal intra-action information, i.e. how action instances are made of sub-actions, and inter-action information, i.e. one specific sub-action may commonly appear in various actions.
****

Graph-Guided Architecture Search for Real-Time Semantic Segmentation
Peiwen Lin, Peng Sun, Guangliang Cheng, Sirui Xie, Xi Li, Jianping Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4203-4212
Designing a lightweight semantic segmentation network often requires researchers to find a trade-off between performance and speed, which is always empirical due to the limited interpretability of neural networks. In order to release researchers from these tedious mechanical trials, we propose a Graph-guided Architecture Search (GAS) pipeline to automatically search real-time semantic segmentation networks. Unlike previous works that use a simplified search space and stack a repeatable cell to form a network, we introduce a novel search mechanism with a new search space where a lightweight model can be effectively explored through the cell-level diversity and latency oriented constraint. Specifically, to produce the cell-level diversity, the cell-sharing constraint is eliminated through the cell-independent manner. Then a graph convolution network (GCN) is seamlessly integrated as a communication mechanism between cells. Finally, a latency-oriented constraint is endowed into the search process to balance the speed and performance. Extensive experiments on Cityscapes and CamVid datasets demonstrate that GAS achieves the new state-of-the-art trade-off between accuracy and speed. In particular, on Cityscapes dataset, GAS achieves the new best performance of 73.5% mIoU with the speed of 108.4 FPS on Titan Xp.
****

Polarized Reflection Removal With Perfect Alignment in the Wild
Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit

ion (CVPR), 2020, pp. 1750-1758
We present a novel formulation to removing reflection from polarized images in t he wild. We first identify the misalignment issues of existing reflection remova l datasets where the collected reflection-free images are not perfectly aligned with input mixed images due to glass refraction. Then we build a new dataset wit h more than 100 types of glass in which obtained transmission images are perfect ly aligned with input mixed images. Second, capitalizing on the special relation ship between reflection and polarized light, we propose a polarized reflection r emoval model with a two-stage architecture. In addition, we design a novel perce ptual NCC loss that can improve the performance of reflection removal and genera l image decomposition tasks. We conduct extensive experiments, and results sugge st that our model outperforms state-of-the-art methods on reflection removal.
****

Blur Aware Calibration of Multi-Focus Plenoptic Camera
Mathieu Labussiere, Celine Teuliere, Frederic Bernardin, Omar Ait-Aider; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 2545-2554
This paper presents a novel calibration algorithm for Multi-Focus Plenoptic Came ras (MFPCs) using raw images only. The design of such cameras is usually complex and relies on precise placement of optic elements. Several calibration procedur es have been proposed to retrieve the camera parameters but relying on simplifie d models, reconstructed images to extract features, or multiple calibrations whe n several types of micro-lens are used. Considering blur information, we propose a new Blur Aware Plenoptic (BAP) feature. It is first exploited in a pre-calibr ation step that retrieves initial camera parameters, and secondly to express a n ew cost function for our single optimization process. The effectiveness of our c alibration method is validated by quantitative and qualitative experiments.
****

Single-Shot Monocular RGB-D Imaging Using Uneven Double Refraction
Andreas Meuleman, Seung-Hwan Baek, Felix Heide, Min H. Kim; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 2465-2474
Cameras that capture color and depth information have become an essential imagin g modality for applications in robotics, autonomous driving, virtual, and augmen ted reality. Existing RGB-D cameras rely on multiple sensors or active illuminat ion with specialized sensors. In this work, we propose a method for monocular si ngle-shot RGB-D imaging. Instead of learning depth from single-image depth cues, we revisit double-refraction imaging using a birefractive medium, measuring dep th as the displacement of differently refracted images superimposed in a single capture. However, existing double-refraction methods are orders of magnitudes to o slow to be used in real-time applications, e.g., in robotics, and provide only inaccurate depth due to correspondence ambiguity in double reflection. We resol ve this ambiguity optically by leveraging the orthogonality of the two linearly polarized rays in double refraction -- introducing uneven double refraction by a dding a linear polarizer to the birefractive medium. Doing so makes it possible to develop a real-time method for reconstructing sparse depth and color simultan eously in real-time. We validate the proposed method, both synthetically and exp erimentally, and demonstrate 3D object detection and photographic applications.
****

Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Laten t Embeddings
Paul Bergmann, Michael Fauser, David Sattlegger, Carsten Steger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 20, pp. 4183-4192
We introduce a powerful student-teacher framework for the challenging problem of unsupervised anomaly detection and pixel-precise anomaly segmentation in high-r esolution images. Student networks are trained to regress the output of a descri ptive teacher network that was pretrained on a large dataset of patches from nat ural images. This circumvents the need for prior data annotation. Anomalies are detected when the outputs of the student networks differ from that of the teache

r network. This happens when they fail to generalize outside the manifold of ano maly-free training data. The intrinsic uncertainty in the student networks is us ed as an additional scoring function that indicates anomalies. We compare our me thod to a large number of existing deep learning based methods for unsupervised anomaly detection. Our experiments demonstrate improvements over state-of-the-ar t methods on a number of real-world datasets, including the recently introduced MVTec Anomaly Detection dataset that was specifically designed to benchmark anom aly segmentation algorithms.
****

Learning to Restore Low-Light Images via Decomposition-and-Enhancement
Ke Xu, Xin Yang, Baocai Yin, Rynson W.H. Lau; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2281-2290
Low-light images typically suffer from two problems. First, they have low visibi lity (i.e., small pixel values). Second, noise becomes significant and disrupts the image content, due to low signal-to-noise ratio. Most existing lowlight imag e enhancement methods, however, learn from noise-negligible datasets. They rely on users having good photographic skills in taking images with low noise. Unfort unately, this is not the case for majority of the low-light images. While concur rently enhancing a low-light image and removing its noise is ill-posed, we obser ve that noise exhibits different levels of contrast in different frequency layer s, and it is much easier to detect noise in the lowfrequency layer than in the h igh one. Inspired by this observation, we propose a frequency-based decompositio nand- enhancement model for low-light image enhancement. Based on this model, we present a novel network that first learns to recover image objects in the low-f requency layer and then enhances high-frequency details based on the recovered i mage objects. In addition, we have prepared a new low-light image dataset with r eal noise to facilitate learning. Finally, we have conducted extensive experimen ts to show that the proposed method outperforms state-of-the-art approaches in e nhancing practical noisy low-light images.
****

HRank: Filter Pruning Using High-Rank Feature Map
Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1529-1538
Neural network pruning offers a promising prospect to facilitate deploying deep neural networks on resource-limited devices. However, existing methods are still challenged by the training inefficiency and labor cost in pruning designs, due to missing theoretical guidance of non-salient network components. In this paper , we propose a novel filter pruning method by exploring the High Rank of feature maps (HRank). Our HRank is inspired by the discovery that the average rank of m ultiple feature maps generated by a single filter is always the same, regardless of the number of image batches CNNs receive. Based on HRank, we develop a metho d that is mathematically formulated to prune filters with low-rank feature maps. The principle behind our pruning is that low-rank feature maps contain less inf ormation, and thus pruned results can be easily reproduced. Besides, we experime ntally show that weights with high-rank feature maps contain more important info rmation, such that even when a portion is not updated, very little damage would be done to the model performance. Without introducing any additional constraints , HRank leads to significant improvements over the state-of-the-arts in terms of FLOPs and parameters reduction, with similar accuracies. For example, with ResN et-110, we achieve a 58.2%-FLOPs reduction by removing 59.2% of the parameters, with only a small loss of 0.14% in top-1 accuracy on CIFAR-10. With Res-50, we a chieve a 43.8%-FLOPs reduction by removing 36.7% of the parameters, with only a loss of 1.17% in the top-1 accuracy on ImageNet. The codes can be available at h ttps://github.com/lmbxmu/HRank.
****

AANet: Adaptive Aggregation Network for Efficient Stereo Matching
Haofei Xu, Juyong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2020, pp. 1959-1968
Despite the remarkable progress made by learning based stereo matching algorithm

s, one key challenge remains unsolved. Current state-of-the-art stereo models ar e mostly based on costly 3D convolutions, the cubic computational complexity and high memory consumption make it quite expensive to deploy in real-world applica tions. In this paper, we aim at completely replacing the commonly used 3D convol utions to achieve fast inference speed while maintaining comparable accuracy. To this end, we first propose a sparse points based intra-scale cost aggregation m ethod to alleviate the well-known edge-fattening issue at disparity discontinuit ies. Further, we approximate traditional cross-scale cost aggregation algorithm with neural network layers to handle large textureless regions. Both modules are simple, lightweight, and complementary, leading to an effective and efficient a rchitecture for cost aggregation. With these two modules, we can not only signif icantly speed up existing top-performing models (e.g., 41x than GC-Net, 4x than PSMNet and 38x than GA-Net), but also improve the performance of fast stereo mod els (e.g., StereoNet). We also achieve competitive results on Scene Flow and KIT TI datasets while running at 62ms, demonstrating the versatility and high effici ency of the proposed method. Our full framework is available at https://github.c om/haofeixu/aanet.

****

Unbiased Scene Graph Generation From Biased Training

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, Hanwang Zhang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2020, pp. 3716-3725

Today's scene graph generation (SGG) task is still far from practical, mainly du e to the severe training bias, e.g., collapsing diverse "human walk on / sit on / lay on beach" into "human on beach". Given such SGG, the down-stream tasks suc h as VQA can hardly infer better scene structures than merely a bag of objects. However, debiasing in SGG is not trivial because traditional debiasing methods c annot distinguish between the good and bad bias, e.g., good context prior (e.g., "person read book" rather than "eat") and bad long-tailed bias (e.g., "near" do minating "behind / in front of"). In this paper, we present a novel SGG framewor k based on causal inference but not the conventional likelihood. We first build a causal graph for SGG, and perform traditional biased training with the graph. Then, we propose to draw the counterfactual causality from the trained graph to infer the effect from the bad bias, which should be removed. In particular, we u se Total Direct Effect (TDE) as the proposed final predicate score for unbiased SGG. Note that our framework is agnostic to any SGG model and thus can be widely applied in the community who seeks unbiased predictions. By using the proposed Scene Graph Diagnosis toolkit on the SGG benchmark Visual Genome and several pre vailing models, we observed significant improvements over the previous state-of- the-art methods.

****

A Semi-Supervised Assessor of Neural Architectures

Yehui Tang, Yunhe Wang, Yixing Xu, Hanting Chen, Boxin Shi, Chao Xu, Chunj ing Xu, Qi Tian, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1810-1819

Neural architecture search (NAS) aims to automatically design deep neural networ ks of satisfactory performance. Wherein, architecture performance predictor is c ritical to efficiently value an intermediate neural architecture. But for the tr aining of this predictor, a number of neural architectures and their correspondi ng real performance often have to be collected. In contrast with classical perfo rmance predictor optimized in a fully supervised way, this paper suggests a semi -supervised assessor of neural architectures. We employ an auto-encoder to disco ver meaningful representations of neural architectures. Taking each neural archi tecture as an individual instance in the search space, we construct a graph to c apture their intrinsic similarities, where both labeled and unlabeled architectu res are involved. A graph convolutional neural network is introduced to predict the performance of architectures based on the learned representations and their relation modeled by the graph. Extensive experimental results on the NAS-Benchma rk-101 dataset demonstrated that our method is able to make a significant reduct ion on the required fully trained architectures for finding efficient architectu

res.
****
Proxy Anchor Loss for Deep Metric Learning
Sungyeon Kim, Dongwon Kim, Minsu Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3238-3247
Existing metric learning losses can be categorized into two classes: pair-based and proxy-based losses. The former class can leverage fine-grained semantic relations between data points, but slows convergence in general due to its high training complexity. In contrast, the latter class enables fast and reliable convergence, but cannot consider the rich data-to-data relations. This paper presents a new proxy-based loss that takes advantages of both pair- and proxy-based methods and overcomes their limitations. Thanks to the use of proxies, our loss boosts the speed of convergence and is robust against noisy labels and outliers. At the same time, it allows embedding vectors of data to interact with each other in its gradients to exploit data-to-data relations. Our method is evaluated on four public benchmarks, where a standard network trained with our loss achieves state-of-the-art performance and most quickly converges.
****
Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs
Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, Heiko Hoffmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 301-310
The unprecedented success of deep neural networks in many applications has made these networks a prime target for adversarial exploitation. In this paper, we introduce a benchmark technique for detecting backdoor attacks (aka Trojan attacks) on deep convolutional neural networks (CNNs). We introduce the concept of Universal Litmus Patterns (ULPs), which enable one to reveal backdoor attacks by feeding these universal patterns to the network and analyzing the output (i.e., classifying the network as `clean' or `corrupted'). This detection is fast because it requires only a few forward passes through a CNN. We demonstrate the effectiveness of ULPs for detecting backdoor attacks on thousands of networks with different architectures trained on four benchmark datasets, namely the German Traffic Sign Recognition Benchmark (GTSRB), MNIST, CIFAR10, and Tiny-ImageNet. The codes and train/test models for this paper can be found here: https://umbcvision.github.io/Universal-Litmus-Patterns/.
****
PQ-NET: A Generative Part Seq2Seq Network for 3D Shapes
Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, Baoquan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 829-838
We introduce PQ-NET, a deep neural network which represents and generates 3D shapes via sequential part assembly. The input to our network is a 3D shape segmented into parts, where each part is first encoded into a feature representation using a part autoencoder. The core component of PQ-NET is a sequence-to-sequence or Seq2Seq autoencoder which encodes a sequence of part features into a latent vector of fixed size, and the decoder reconstructs the 3D shape, one part at a time, resulting in a sequential assembly. The latent space formed by the Seq2Seq encoder encodes both part structure and fine part geometry. The decoder can be adapted to perform several generative tasks including shape autoencoding, interpolation, novel shape generation, and single-view 3D reconstruction, where the generated shapes are all composed of meaningful parts.
****
Extreme Relative Pose Network Under Hybrid Representations
Zhenpei Yang, Siming Yan, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2455-2464
In this paper, we introduce a novel RGB-D based relative pose estimation approach that is suitable for small-overlapping or non-overlapping scans and can output multiple relative poses. Our method performs scene completion and matches the completed scans. However, instead of using a fixed representation for completion,

the key idea is to utilize hybrid representations that combine 360-image, 2D image-based layout, and planar patches. This approach offers adaptively feature representations for relative pose estimation. Besides, we introduce a global-2-local matching procedure, which utilizes initial relative poses obtained during the global phase to detect and then integrate geometric relations for pose refinement. Experimental results justify the potential of this approach across a wide range of benchmark datasets. For example, on ScanNet, the rotation translation errors of the top-1/top-5 predictions of our approach are 28.6^ \circ /0.90m and 16.8^ \circ /0.76m, respectively. Our approach also considerably boosts the performance of multi-scan reconstruction in few-view reconstruction settings.
****

Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline
Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1651-1660
Recovering a high dynamic range (HDR) image from a single low dynamic range (LDR) input image is challenging due to missing details in under-/over-exposed regions caused by quantization and saturation of camera sensors. In contrast to existing learning-based methods, our core idea is to incorporate the domain knowledge of the LDR image formation pipeline into our model. We model the HDR-to-LDR image formation pipeline as the (1) dynamic range clipping, (2) non-linear mapping from a camera response function, and (3) quantization. We then propose to learn three specialized CNNs to reverse these steps. By decomposing the problem into specific sub-tasks, we impose effective physical constraints to facilitate the training of individual sub-networks. Finally, we jointly fine-tune the entire model end-to-end to reduce error accumulation. With extensive quantitative and qualitative experiments on diverse image datasets, we demonstrate that the proposed method performs favorably against state-of-the-art single-image HDR reconstruction algorithms.
****

Learning to Observe: Approximating Human Perceptual Thresholds for Detection of Suprathreshold Image Transformations
Alan Dolhasz, Carlo Harvey, Ian Williams; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4797-4807
Many tasks in computer vision are often calibrated and evaluated relative to human perception. In this paper, we propose to directly approximate the perceptual function performed by human observers completing a visual detection task. Specifically, we present a novel methodology for learning to detect image transformations visible to human observers through approximating perceptual thresholds. To do this, we carry out a subjective two-alternative forced-choice study to estimate perceptual thresholds of human observers detecting local exposure shifts in images. We then leverage transformation equivariant representation learning to overcome issues of limited perceptual data. This representation is then used to train a dense convolutional classifier capable of detecting local suprathreshold exposure shifts - a distortion common to image composites. In this context, our model can approximate perceptual thresholds with an average error of 0.1148 exposure stops between empirical and predicted thresholds. It can also be trained to detect a range of different local transformations.
****

MEBOW: Monocular Estimation of Body Orientation in the Wild
Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzra, Zhuo Deng, Bilan Liu, James Z. Wang, Cheng-hao Kuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3451-3461
Body orientation estimation provides crucial visual cues in many applications, including robotics and autonomous driving. It is particularly desirable when 3-D pose estimation is difficult to infer due to poor image resolution, occlusion or indistinguishable body parts. We present COCO-MEBOW (Monocular Estimation of Body Orientation in the Wild), a new large-scale dataset for orientation estimation from a single in-the-wild image. The body-orientation labels for around 130K h

uman bodies within 55K images from the COCO dataset have been collected using an efficient and high-precision annotation pipeline. We also validated the benefits of the dataset. First, we show that our dataset can substantially improve the performance and the robustness of a human body orientation estimation model, the development of which was previously limited by the scale and diversity of the available training data. Additionally, we present a novel triple-source solution for 3-D human pose estimation, where 3-D pose labels, 2-D pose labels, and our body-orientation labels are all used in joint training. Our model significantly outperforms state-of-the-art dual-source solutions for monocular 3-D human pose estimation, where training only uses 3-D pose labels and 2-D pose labels. This substantiates an important advantage of MEBOW for 3-D human pose estimation, which is particularly appealing because the per-instance labeling cost for body orientations is far less than that for 3-D poses. The work demonstrates high potential of MEBOW in addressing real-world challenges involving understanding human behaviors. Further information of this work is available at https://chenyanwu.github.io/MEBOW/.

****

Depth Sensing Beyond LiDAR Range

Kai Zhang, Jiaxin Xie, Noah Snavely, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1692-1700

Depth sensing is a critical component of autonomous driving technologies, but today's LiDAR- or stereo camera- based solutions have limited range. We seek to increase the maximum range of self-driving vehicles' depth perception modules for the sake of better safety. To that end, we propose a novel three-camera system that utilizes small field of view cameras. Our system, along with our novel algorithm for computing metric depth, does not require full pre-calibration and can output dense depth maps with practically acceptable accuracy for scenes and objects at long distances not well covered by most commercial LiDARs.

****

BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1790-1799

While deep learning has recently achieved great success on multi-view stereo (MVS), limited training data makes the trained model hard to be generalized to unseen scenarios. Compared with other computer vision tasks, it is rather difficult to collect a large-scale MVS dataset as it requires expensive active scanners and labor-intensive process to obtain ground truth 3D structures. In this paper, we introduce BlendedMVS, a novel large-scale dataset, to provide sufficient training ground truth for learning-based MVS. To create the dataset, we apply a 3D reconstruction pipeline to recover high-quality textured meshes from images of well-selected scenes. Then, we render these mesh models to color images and depth maps. To introduce the ambient lighting information during training, the rendered color images are further blended with the input images to generate the training input. Our dataset contains over 17k high-resolution images covering a variety of scenes, including cities, architectures, sculptures and small objects. Extensive experiments demonstrate that BlendedMVS endows the trained model with significantly better generalization ability compared with other MVS datasets. The dataset and pretrained models are available at https://github.com/YoYo000/BlendedMVS.

****

Learning to Detect Important People in Unlabelled Images for Semi-Supervised Important People Detection

Fa-Ting Hong, Wei-Hong Li, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4146-4154

Important people detection is to automatically detect the individuals who play the most important roles in a social event image, which requires the designed model to understand a high-level pattern. However, existing methods rely heavily on supervised learning using large quantities of annotated image samples, which ar

e more costly to collect for important people detection than for individual enti ty recognition (i.e., object recognition). To overcome this problem, we propose learning important people detection on partially annotated images. Our approach iteratively learns to assign pseudo-labels to individuals in un-annotated images and learns to update the important people detection model based on data with bo th labels and pseudo-labels. To alleviate the pseudo-labelling imbalance problem , we introduce a ranking strategy for pseudo-label estimation, and also introduc e two weighting strategies: one for weighting the confidence that individuals ar e important people to strengthen the learning on important people and the other for neglecting noisy unlabelled images (i.e., images without any important peopl e). We have collected two large-scale datasets for evaluation. The extensive exp erimental results clearly confirm the efficacy of our method attained by leverag ing unlabelled images for improving the performance of important people detectio n.

****

Fixed-Point Back-Propagation Training

Xishan Zhang, Shaoli Liu, Rui Zhang, Chang Liu, Di Huang, Shiyi Zhou, Jiam ing Guo, Qi Guo, Zidong Du, Tian Zhi, Yunji Chen; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2330-2 338

Recent emerged quantization technique (i.e., using low bit-width fixed-point dat a instead of high bit-width floating-point data) has been applied to inference o f deep neural networks for fast and efficient execution. However, directly apply ing quantization in training can cause significant accuracy loss, thus remaining an open challenge. In this paper, we propose a novel training approach, which a pplies a layer-wise precision-adaptive quantization in deep neural networks. The new training approach leverages our key insight that the degradation of trainin g accuracy is attributed to the dramatic change of data distribution. Therefore, by keeping the data distribution stable through a layer-wise precision-adaptive quantization, we are able to directly train deep neural networks using low bit- width fixed-point data and achieve guaranteed accuracy, without changing hyper p arameters. Experimental results on a wide variety of network architectures (e.g. , convolution and recurrent networks) and applications (e.g., image classificati on, object detection, segmentation and machine translation) show that the propos ed approach can train these neural networks with negligible accuracy losses (-1. 40%-1.3%, 0.02% on average), and speed up training by 252% on a state-of-the-art Intel CPU.

****

Zero-Assignment Constraint for Graph Matching With Outliers

Fudong Wang, Nan Xue, Jin-Gang Yu, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3033-304 2

Graph matching (GM), as a longstanding problem in computer vision and pattern re cognition, still suffers from numerous cluttered outliers in practical applicati ons. To address this issue, we present the zero-assignment constraint (ZAC) for approaching the graph matching problem in the presence of outliers. The underlyi ng idea is to suppress the matchings of outliers by assigning zero-valued vector s to the potential outliers in the obtained optimal correspondence matrix. We pr ovide elaborate theoretical analysis to the problem, i.e., GM with ZAC, and figu re out that the GM problem with and without outliers are intrinsically different , which enables us to put forward a sufficient condition to construct valid and reasonable objective function. Consequently, we design an efficient outlier-robu st algorithm to significantly reduce the incorrect or redundant matchings caused by numerous outliers. Extensive experiments demonstrate that our method can ach ieve the state-of-the-art performance in terms of accuracy and efficiency, espec ially in the presence of numerous outliers.

****

AvatarMe: Realistically Renderable 3D Facial Reconstruction "In-the-Wild"

Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Va sileios Triantafyllou, Abhijeet Ghosh, Stefanos Zafeiriou; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 760-769
Over the last years, with the advent of Generative Adversarial Networks (GANs), many face analysis tasks have accomplished astounding performance, with applications including, but not limited to, face generation and 3D face reconstruction from a single "in-the-wild" image. Nevertheless, to the best of our knowledge, there is no method which can produce high-resolution photorealistic 3D faces from "in-the-wild" images and this can be attributed to the: (a) scarcity of available data for training, and (b) lack of robust methodologies that can successfully be applied on very high-resolution data. In this paper, we introduce AvatarMe, the first method that is able to reconstruct photorealistic 3D faces from a single "in-the-wild" image with an increasing level of detail. To achieve this, we capture a large dataset of facial shape and reflectance and build on a state-of-the-art 3D texture and shape reconstruction method and successively refine its results, while generating the per-pixel diffuse and specular components that are required for realistic rendering. As we demonstrate in a series of qualitative and quantitative experiments, AvatarMe outperforms the existing arts by a significant margin and reconstructs authentic, 4K by 6K-resolution 3D faces from a single low-resolution image that, for the first time, bridges the uncanny valley.
****

Generalized Product Quantization Network for Semi-Supervised Image Retrieval
Young Kyun Jang, Nam Ik Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3420-3429
Image retrieval methods that employ hashing or vector quantization have achieved great success by taking advantage of deep learning. However, these approaches do not meet expectations unless expensive label information is sufficient. To resolve this issue, we propose the first quantization-based semi-supervised image retrieval scheme: Generalized Product Quantization (GPQ) network. We design a novel metric learning strategy that preserves semantic similarity between labeled data, and employ entropy regularization term to fully exploit inherent potentials of unlabeled data. Our solution increases the generalization capacity of the quantization network, which allows overcoming previous limitations in the retrieval community. Extensive experimental results demonstrate that GPQ yields state-of-the-art performance on large-scale real image benchmark datasets.
****

Extremely Dense Point Correspondences Using a Learned Feature Descriptor
Xingtong Liu, Yiping Zheng, Benjamin Killeen, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, Mathias Unberath; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4847-4856
High-quality 3D reconstructions from endoscopy video play an important role in many clinical applications, including surgical navigation where they enable direct video-CT registration. While many methods exist for general multi-view 3D reconstruction, these methods often fail to deliver satisfactory performance on endoscopic video. Part of the reason is that local descriptors that establish pair-wise point correspondences, and thus drive reconstruction, struggle when confronted with the texture-scarce surface of anatomy. Learning-based dense descriptors usually have larger receptive fields enabling the encoding of global information, which can be used to disambiguate matches. In this work, we present an effective self-supervised training scheme and novel loss design for dense descriptor learning. In direct comparison to recent local and dense descriptors on an in-house sinus endoscopy dataset, we demonstrate that our proposed dense descriptor can generalize to unseen patients and scopes, thereby largely improving the performance of Structure from Motion (SfM) in terms of model density and completeness. We also evaluate our method on a public dense optical flow dataset and a small-scale SfM public dataset to further demonstrate the effectiveness and generality of our method. The source code is available at https://github.com/lppllppl920/DenseDescriptorLearning-Pytorch.
****

Gate-Shift Networks for Video Action Recognition
Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz; Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1102-1111
Deep 3D CNNs for video action recognition are designed to learn powerful representations in the joint spatio-temporal feature space. In practice however, because of the large number of parameters and computations involved, they may under-perform in the lack of sufficiently large datasets for training them at scale. In this paper we introduce spatial gating in spatial-temporal decomposition of 3D kernels. We implement this concept with Gate-Shift Module (GSM). GSM is lightweight and turns a 2D-CNN into a highly efficient spatio-temporal feature extractor. With GSM plugged in, a 2D-CNN learns to adaptively route features through time and combine them, at almost no additional parameters and computational overhead. We perform an extensive evaluation of the proposed module to study its effectiveness in video action recognition, achieving state-of-the-art results on Something Something-V1 and Diving48 datasets, and obtaining competitive results on EPIC-Kitchens with far less model complexity.
****

CRNet: Cross-Reference Networks for Few-Shot Segmentation
Weide Liu, Chi Zhang, Guosheng Lin, Fayao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4165-4173
Over the past few years, state-of-the-art image segmentation algorithms are based on deep convolutional neural networks. To render a deep network with the ability to understand a concept, humans need to collect a large amount of pixel-level annotated data to train the models, which is time-consuming and tedious. Recently, few-shot segmentation is proposed to solve this problem. Few-shot segmentation aims to learn a segmentation model that can be generalized to novel classes with only a few training images. In this paper, we propose a cross-reference network (CRNet) for few-shot segmentation. Unlike previous works which only predict the mask in the query image, our proposed model concurrently makes predictions for both the support image and the query image. With a cross-reference mechanism, our network can better find the co-occurrent objects in two images, thus helping the few-shot segmentation task. We also develop a mask refinement module to recurrently refine the prediction of the foreground regions. For the k-shot learning, we propose to finetune parts of networks to take advantage of multiple labeled support images. Experiments on the PASCAL VOC 2012 dataset show that our network achieves state-of-the-art performance.
****

Space-Time-Aware Multi-Resolution Video Enhancement
Muhammad Haris, Greg Shakhnarovich, Norimichi Ukita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2859-2868
We consider the problem of space-time super-resolution (ST-SR): increasing spatial resolution of video frames and simultaneously interpolating frames to increase the frame rate. Modern approaches handle these axes one at a time. In contrast, our proposed model called STARnet super-resolves jointly in space and time. This allows us to leverage mutually informative relationships between time and space: higher resolution can provide more detailed information about motion, and higher frame-rate can provide better pixel alignment. The components of our model that generate latent low- and high-resolution representations during ST-SR can be used to finetune a specialized mechanism for just spatial or just temporal super-resolution. Experimental results demonstrate that STARnet improves the performances of space-time, spatial, and temporal video super-resolution by substantial margins on publicly available datasets.
****

METAL: Minimum Effort Temporal Activity Localization in Untrimmed Videos
Da Zhang, Xiyang Dai, Yuan-Fang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3882-3892
Existing Temporal Activity Localization (TAL) methods largely adopt strong supervision for model training, which requires (1) vast amounts of untrimmed videos per each activity category and (2) accurate segment-level boundary annotations (start time and end time) for every instance. This poses a critical restriction to

the current methods in practical scenarios where not only segment-level annotations are expensive to obtain, but many activity categories are also rare and unobserved during training. Therefore, Can we learn a TAL model under weak supervision that can localize unseen activity classes? To address this scenario, we define a novel example-based TAL problem called Minimum Effort Temporal Activity Localization (METAL): Given only a few examples, the goal is to find the occurrences of semantically-related segments in an untrimmed video sequence while model training is only supervised by the video-level annotation. Towards this objective, we propose a novel Similarity Pyramid Network (SPN) that adopts the few-shot learning technique of Relation Network and directly encodes hierarchical multi-scale correlations, which we learn by optimizing two complimentary loss functions in an end-to-end manner. We evaluate the SPN on the THUMOS'14 and ActivityNet datasets, of which we rearrange the videos to fit the METAL setup. Results show that our SPN achieves performance superior or competitive to state-of-the-art approaches with stronger supervision.
****

Image Demoireing with Learnable Bandpass Filters
Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, Ales Leonardis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3636-3645
Image demoireing is a multi-faceted image restoration task involving both texture and color restoration. In this paper, we propose a novel multiscale bandpass convolutional neural network (MBCNN) to address this problem. As an end-to-end solution, MBCNN respectively solves the two sub-problems. For texture restoration, we propose a learnable bandpass filter (LBF) to learn the frequency prior for moire texture removal. For color restoration, we propose a two-step tone mapping strategy, which first applies a global tone mapping to correct for a global color shift, then performs local fine tuning of the color per pixel. Through an ablation study, we demonstrate the effectiveness of the different components of MBCNN. Experimental results on two public datasets show that our method outperforms state-of-the-art methods by a large margin (more than 2dB in terms of PSNR).
****

Active Vision for Early Recognition of Human Actions
Boyu Wang, Lihan Huang, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1081-1091
We propose a method for early recognition of human actions, one that can take advantages of multiple cameras while satisfying the constraints due to limited communication bandwidth and processing power. Our method considers multiple cameras, and at each time step, it will decide the best camera to use so that a confident recognition decision can be reached as soon as possible. We formulate the camera selection problem as a sequential decision process, and learn a view selection policy based on reinforcement learning. We also develop a novel recurrent neural network architecture to account for the unobserved video frames and the irregular intervals between the observed frames. Experiments on three datasets demonstrate the effectiveness of our approach for early recognition of human actions.
****

Weakly-Supervised Action Localization by Generative Attention Modeling
Baifeng Shi, Qi Dai, Yadong Mu, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1009-1019
Weakly-supervised temporal action localization is a problem of learning an action localization model with only video-level action labeling available. The general framework largely relies on the classification activation, which employs an attention model to identify the action-related frames and then categorizes them into different classes. Such method results in the action-context confusion issue: context frames near action clips tend to be recognized as action frames themselves, since they are closely related to the specific classes. To solve the problem, in this paper we propose to model the class-agnostic frame-wise probability conditioned on the frame attention using conditional Variational Auto-Encoder (VAE). With the observation that the context exhibits notable difference from the action at representation level, a probabilistic model, i.e., conditional VAE, is

learned to model the likelihood of each frame given the attention. By maximizing the conditional probability with respect to the attention, the action and non-action frames are well separated. Experiments on THUMOS14 and ActivityNet1.2 demonstrate advantage of our method and effectiveness in handling action-context confusion problem. Code is now available on GitHub.

****

Unsupervised Domain Adaptation With Hierarchical Gradient Synchronization

Lanqing Hu, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4043-4052

Domain adaptation attempts to boost the performance on a target domain by borrowing knowledge from a well established source domain. To handle the distribution gap between two domains, the prominent approaches endeavor to extract domain-invariant features. It is known that after a perfect domain alignment the domain-invariant representations of two domains should share the same characteristics from perspective of the overview and also any local piece. Inspired by this, we propose a novel method called Hierarchical Gradient Synchronization to model the synchronization relationship among the local distribution pieces and global distribution, aiming for more precise domain-invariant features. Specifically, the hierarchical domain alignments including class-wise alignment, group-wise alignment and global alignment are first constructed. Then, these three types of alignment are constrained to be consistent to ensure better structure preservation. As a result, the obtained features are domain invariant and intrinsically structure preserved. As evaluated on extensive domain adaptation tasks, our proposed method achieves state-of-the-art classification performance on both vanilla unsupervised domain adaptation and partial domain adaptation.

****

Learning Unsupervised Hierarchical Part Decomposition of 3D Objects From a Single RGB Image

Despoina Paschalidou, Luc Van Gool, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1060-1070

Humans perceive the 3D world as a set of distinct objects that are characterized by various low-level (geometry, reflectance) and high-level (connectivity, adjacency, symmetry) properties. Recent methods based on convolutional neural networks (CNNs) demonstrated impressive progress in 3D reconstruction, even when using a single 2D image as input. However, the majority of these methods focuses on recovering the local 3D geometry of an object without considering its part-based decomposition or relations between parts. We address this challenging problem by proposing a novel formulation that allows to jointly recover the geometry of a 3D object as a set of primitives as well as their latent hierarchical structure without part-level supervision. Our model recovers the higher level structural decomposition of various objects in the form of a binary tree of primitives, where simple parts are represented with fewer primitives and more complex parts are modeled with more components. Our experiments on the ShapeNet and D-FAUST datasets demonstrate that considering the organization of parts indeed facilitates reasoning about 3D geometry.

****

Light Field Spatial Super-Resolution via Deep Combinatorial Geometry Embedding and Structural Consistency Regularization

Jing Jin, Junhui Hou, Jie Chen, Sam Kwong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2260-2269

Light field (LF) images acquired by hand-held devices usually suffer from low spatial resolution as the limited sampling resources have to be shared with the angular dimension. LF spatial super-resolution (SR) thus becomes an indispensable part of the LF camera processing pipeline. The high-dimensionality characteristic and complex geometrical structure of LF images makes the problem more challenging than traditional single-image SR. The performance of existing methods are still limited as they fail to thoroughly explore the coherence among LF views and are insufficient in accurately preserving the parallax structure of the scene. I

n this paper, we propose a novel learning-based LF spatial SR framework, in whic h each view of an LF image is first individually super-resolved by exploring the complementary information among views with combinatorial geometry embedding. Fo r accurate preservation of the parallax structure among the reconstructed views, a regularization network trained over a structure-aware loss function is subseq uently appended to enforce correct parallax relationships over the intermediate estimation. Our proposed approach is evaluated over datasets with a large number of testing images including both synthetic and real-world scenes. Experimental results demonstrate the advantage of our approach over state-of-the-art methods, i.e., our method not only improves the average PSNR by more than 1.0 dB but als o preserves more accurate parallax details, at a lower computation cost.
****

Auto-Encoding Twin-Bottleneck Hashing
Yuming Shen, Jie Qin, Jiaxin Chen, Mengyang Yu, Li Liu, Fan Zhu, Fumin She n, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2020, pp. 2818-2827
Conventional unsupervised hashing methods usually take advantage of similarity g raphs, which are either pre-computed in the high-dimensional space or obtained f rom random anchor points. On the one hand, existing methods uncouple the procedu res of hash function learning and graph construction. On the other hand, graphs empirically built upon original data could introduce biased prior knowledge of d ata relevance, leading to sub-optimal retrieval performance. In this paper, we t ackle the above problems by proposing an efficient and adaptive code-driven grap h, which is updated by decoding in the context of an auto-encoder. Specifically, we introduce into our framework twin bottlenecks (i.e., latent variables) that exchange crucial information collaboratively. One bottleneck (i.e., binary codes) conveys the high-level intrinsic data structure captured by the code-driven gr aph to the other (i.e., continuous variables for low-level detail information), which in turn propagates the updated network feedback for the encoder to learn m ore discriminative binary codes. The auto-encoding learning objective literally rewards the code-driven graph to learn an optimal encoder. Moreover, the propose d model can be simply optimized by gradient descent without violating the binary constraints. Experiments on benchmarked datasets clearly show the superiority o f our framework over the state-of-the-art hashing methods. Our source code can b e found at https://github.com/ymcidence/TBH.
****

Learning to Super Resolve Intensity Images From Events
S. Mohammad Mostafavi I., Jonghyun Choi, Kuk-Jin Yoon; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 276 8-2776
An event camera detects per-pixel intensity difference and produces asynchronous event stream with low latency, high dynamic range, and low power consumption. A s a trade-off, the event camera has low spatial resolution. We propose an end-to -end network to reconstruct high resolution, high dynamic range (HDR) images dir ectly from the event stream. We evaluate our algorithm on both simulated and rea l-world sequences and verify that it captures fine details of a scene and outper forms the combination of the state-of-the-art event to image algorithms with the state-of-the-art super resolution schemes in many quantitative measures by larg e margins. We further extend our method by using the active sensor pixel (APS) f rames or reconstructing images iteratively.
****

FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D F ace Prediction
Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 601-610
In this paper, we present a large-scale detailed 3D face dataset, FaceScape, and propose a novel algorithm that is able to predict elaborate riggable 3D face mo dels from a single image input. FaceScape dataset provides 18,760 textured 3D fa ces, captured from 938 subjects and each with 20 specific expressions. The 3D mo

dels contain the pore-level facial geometry that is also processed to be topolog ically uniformed. These fine 3D facial models can be represented as a 3D morphab le model for rough shapes and displacement maps for detailed geometry. Taking ad vantage of the large-scale and high-accuracy dataset, a novel algorithm is furth er proposed to learn the expression-specific dynamic details using a deep neural network. The learned relationship serves as the foundation of our 3D face predi ction system from a single image input. Different than the previous methods, our predicted 3D models are riggable with highly detailed geometry under different expressions. The unprecedented dataset and code will be released to public for r esearch purpose.
****

Cross-View Tracking for Multi-Human 3D Pose Estimation at Over 100 FPS
Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, Shuang Liu; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 3279-3288
Estimating 3D poses of multiple humans in real-time is a classic but still chall enging task in computer vision. Its major difficulty lies in the ambiguity in cr oss-view association of 2D poses and the huge state space when there are multipl e people in multiple views. In this paper, we present a novel solution for multi -human 3D pose estimation from multiple calibrated camera views. It takes 2D pos es in different camera coordinates as inputs and aims for the accurate 3D poses in the global coordinate. Unlike previous methods that associate 2D poses among all pairs of views from scratch at every frame, we exploit the temporal consiste ncy in videos to match the 2D inputs with 3D poses directly in 3-space. More spe cifically, we propose to retain the 3D pose for each person and update them iter atively via the cross-view multi-human tracking. This novel formulation improves both accuracy and efficiency, as we demonstrated on widely-used public datasets . To further verify the scalability of our method, we propose a new large-scale multi-human dataset with 12 to 28 camera views. Without bells and whistles, our solution achieves 154 FPS on 12 cameras and 34 FPS on 28 cameras, indicating its ability to handle large-scale real-world applications. The proposed dataset wil l be released at https://github.com/longcw/crossview_3d_pose_tracking.
****

ADINet: Attribute Driven Incremental Network for Retinal Image Classification
Qier Meng, Satoh Shin'ichi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4033-4042
Retinal diseases encompass a variety of types, including different diseases and severity levels. Training a model with different types of disease is impractical . Dynamically training a model is necessary when a patient with a new disease ap pears. Deep learning techniques have stood out in recent years, but they suffer from catastrophic forgetting, i.e., a dramatic decrease in performance when new training classes appear. We found that keeping the feature distribution of an ol d model helps maintain the performance of incremental learning. In this paper, w e design a framework named "Attribute Driven Incremental Network" (ADINet), a ne w architecture that integrates class label prediction and attribute prediction i nto an incremental learning framework to boost the classification performance. W ith image-level classification, we apply knowledge distillation (KD) to retain t he knowledge of base classes. With attribute prediction, we calculate the weight of each attribute of an image and use these weights for more precise attribute prediction. We designed attribute distillation (AD) loss to retain the informati on of base class attributes as new classes appear. This incremental learning can be performed multiple times with a moderate drop in performance. The results of an experiment on our private retinal fundus image dataset demonstrate that our proposed method outperforms existing state-of-the-art methods. For demonstrating the generalization of our proposed method, we test it on the ImageNet-150K-sub dataset and show good performance.
****

Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spa tial Resolution Remote Sensing Imagery
Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4096-4105
Geospatial object segmentation, as a particular semantic segmentation task, always faces with larger-scale variation, larger intra-class variance of background, and foreground-background imbalance in the high spatial resolution (HSR) remote sensing imagery. However, general semantic segmentation methods mainly focus on scale variation in the natural scene, with inadequate consideration of the other two problems that usually happen in the large area earth observation scene. In this paper, we argue that the problems lie on the lack of foreground modeling and propose a foreground-aware relation network (FarSeg) from the perspectives of relation-based and optimization-based foreground modeling, to alleviate the above two problems. From perspective of relation, FarSeg enhances the discrimination of foreground features via foreground-correlated contexts associated by learning foreground-scene relation. Meanwhile, from perspective of optimization, a foreground-aware optimization is proposed to focus on foreground examples and hard examples of background during training for a balanced optimization. The experimental results obtained using a large scale dataset suggest that the proposed method is superior to the state-of-the-art general semantic segmentation methods and achieves a better trade-off between speed and accuracy.
****

Single-View View Synthesis With Multiplane Images
Richard Tucker, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 551-560
A recent strand of work in view synthesis uses deep learning to generate multiplane images--a camera-centric, layered 3D representation--given two or more input images at known viewpoints. We apply this representation to single-view view synthesis, a problem which is more challenging but has potentially much wider application. Our method learns to predict a multiplane image directly from a single image input, and we introduce scale-invariant view synthesis for supervision, enabling us to train on online video. We show this approach is applicable to several different datasets, that it additionally generates reasonable depth maps, and that it learns to fill in content behind the edges of foreground objects in background layers.
****

Select, Supplement and Focus for RGB-D Saliency Detection
Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3472-3481
Depth data containing a preponderance of discriminative power in location have been proven beneficial for accurate saliency prediction. However, RGB-D saliency detection methods are also negatively influenced by randomly distributed erroneous or missing regions on the depth map or along the object boundaries. This offers the possibility of achieving more effective inference by well designed models. In this paper, we propose a new framework for accurate RGB-D saliency detection taking account of local and global complementarities from two modalities. This is achieved by designing a complimentary interaction model discriminative enough to simultaneously select useful representation from RGB and depth data, and meanwhile to refine the object boundaries. Moreover, we proposed a compensation-aware loss to further process the information not being considered in the complimentary interaction model, leading to improvement of the generalization ability for challenging scenes. Experiments on six public datasets show that our method outperforms18state-of-the-art methods.
****

Towards Discriminability and Diversity: Batch Nuclear-Norm Maximization Under Label Insufficient Situations
Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3941-3950
The learning of the deep networks largely relies on the data with human-annotated labels. In some label insufficient situations, the performance degrades on the

decision boundary with high data density. A common solution is to directly minimize the Shannon Entropy, but the side effect caused by entropy minimization, i.e., reduction of the prediction diversity, is mostly ignored. To address this issue, we reinvestigate the structure of classification output matrix of a randomly selected data batch. We find by theoretical analysis that the prediction discriminability and diversity could be separately measured by the Frobenius-norm and rank of the batch output matrix. Besides, the nuclear-norm is an upperbound of the Frobenius-norm, and a convex approximation of the matrix rank. Accordingly, to improve both discriminability and diversity, we propose Batch Nuclear-norm Maximization (BNM) on the output matrix. BNM could boost the learning under typical label insufficient learning scenarios, such as semi-supervised learning, domain adaptation and open domain recognition. On these tasks, extensive experimental results show that BNM outperforms competitors and works well with existing well-known methods. The code is available at https://github.com/cuishuhao/BNM

****

4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras

Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1324-1333

his paper contributes a novel realtime multi-person motion capture algorithm using multiview video inputs. Due to the heavy occlusions and closely interacting motions in each view, joint optimization on the multiview images and multiple temporal frames is indispensable, which brings up the essential challenge of realtime efficiency. To this end, for the first time, we unify per-view parsing, cross-view matching, and temporal tracking into a single optimization framework, i.e., a 4D association graph that each dimension (image space, viewpoint and time) can be treated equally and simultaneously. To solve the 4D association graph efficiently, we further contribute the idea of 4D limb bundle parsing based on heuristic searching, followed with limb bundle assembling by proposing a bundle Kruskal's algorithm. Our method enables a realtime motion capture system running at 30fps using 5 cameras on a 5-person scene. Benefiting from the unified parsing, matching and tracking constraints, our method is robust to noisy detection due to severe occlusions and close interacting motions, and achieves high-quality online pose reconstruction quality. The proposed method outperforms state-of-the-art methods quantitatively without using high-level appearance information.

****

SDC-Depth: Semantic Divide-and-Conquer Network for Monocular Depth Estimation

Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 541-550

Monocular depth estimation is an ill-posed problem, and as such critically relies on scene priors and semantics. Due to its complexity, we propose a deep neural network model based on a semantic divide-and-conquer approach. Our model decomposes a scene into semantic segments, such as object instances and background stuff classes, and then predicts a scale and shift invariant depth map for each semantic segment in a canonical space. Semantic segments of the same category share the same depth decoder, so the global depth prediction task is decomposed into a series of category-specific ones, which are simpler to learn and easier to generalize to new scene types. Finally, our model stitches each local depth segment by predicting its scale and shift based on the global context of the image. The model is trained end-to-end using a multi-task loss for panoptic segmentation and depth prediction, and is therefore able to leverage large-scale panoptic segmentation datasets to boost its semantic understanding. We validate the effectiveness of our approach and show state-of-the-art performance on three benchmark datasets.

****

Meshlet Priors for 3D Mesh Reconstruction

Abhishek Badki, Orazio Gallo, Jan Kautz, Pradeep Sen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 284

9-2858
Estimating a mesh from an unordered set of sparse, noisy 3D points is a challeng ing problem that requires to carefully select priors. Existing hand-crafted prio rs, such as smoothness regularizers, impose an undesirable trade-off between att enuating noise and preserving local detail. Recent deep-learning approaches prod uce impressive results by learning priors directly from the data. However, the p riors are learned at the object level, which makes these algorithms class-specif ic, and even sensitive to the pose of the object. We introduce meshlets, small p atches of mesh that we use to learn local shape priors. Meshlets act as a dictio nary of local features and thus allow to use learned priors to reconstruct objec t meshes in any pose and from unseen classes, even when the noise is large and t he samples sparse.
****

Transferable, Controllable, and Inconspicuous Adversarial Attacks on Person Re-i dentification With Deep Mis-Ranking
Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 342-351
The success of DNNs has driven the extensive applications of person re-identific ation (ReID) into a new era. However, whether ReID inherits the vulnerability of DNNs remains unexplored. To examine the robustness of ReID systems is rather im portant because the insecurity of ReID systems may cause severe losses, e.g., th e criminals may use the adversarial perturbations to cheat the CCTV systems. In this work, we examine the insecurity of current best-performing ReID models by p roposing a learning-to-mis-rank formulation to perturb the ranking of the system output. As the cross-dataset transferability is crucial in the ReID domain, we also perform a back-box attack by developing a novel multi-stage network archite cture that pyramids the features of different levels to extract general and tran sferable features for the adversarial perturbations. Our method can control the number of malicious pixels by using differentiable multi-shot sampling. To guara ntee the inconspicuousness of the attack, we also propose a new perception loss to achieve better visual quality. Extensive experiments on four of the largest R eID benchmarks (i.e., Market1501, CUHK03, DukeMTMC, and MSMT17) not only show th e effectiveness of our method, but also provides directions of the future improv ement in the robustness of ReID systems. For example, the accuracy of one of the best-performing ReID systems drops sharply from 91.8% to 1.4% after being attac ked by our method. Some attack results are shown in Fig. 1. The code is availabl e at: https://github.com/whj363636/Adversarial-attack-on-Person-ReID-With-Deep-M is-Ranking.
****

More Grounded Image Captioning by Distilling Image-Text Matching Model
Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 20, pp. 4777-4786
Visual attention not only improves the performance of image captioners, but also serves as a visual interpretation to qualitatively measure the caption rational ity and model transparency. Specifically, we expect that a captioner can fix its attentive gaze on the correct objects while generating the corresponding words. This ability is also known as grounded image captioning. However, the grounding accuracy of existing captioners is far from satisfactory.To improve the groundi ng accuracy while retaining the captioning quality, it is expensive to collect t he word-region alignment as strong supervision.To this end, we propose a Part-of -Speech (POS) enhanced image-text matching model (SCAN): POS-SCAN, as the effect ive knowledge distillation for more grounded image captioning. The benefits are two-fold: 1) given a sentence and an image, POS-SCAN can ground the objects more accurately than SCAN; 2) POS-SCAN serves as a word-region alignment regularizat ion for the captioner's visual attention module. By showing benchmark experiment al results, we demonstrate that conventional image captioners equipped with POS-SCAN can significantly improve the grounding accuracy without strong supervision . Last but not the least, we explore the indispensable Self-Critical Sequence Tr

aining (SCST) in the context of grounded image captioning and show that the imag
e-text matching score can serve as a reward for more grounded captioning.
****

Leveraging Photometric Consistency Over Time for Sparsely Supervised Hand-Object
 Reconstruction

Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, Corde
lia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2020, pp. 571-580

Modeling hand-object manipulations is essential for understanding how humans int
eract with their environment. While of practical importance, estimating the pose
 of hands and objects during interactions is challenging due to the large mutual
 occlusions that occur during manipulation. Recent efforts have been directed to
wards fully-supervised methods that require large amounts of labeled training sa
mples. Collecting 3D ground-truth data for hand-object interactions, however, is
 costly, tedious, and error-prone. To overcome this challenge we present a metho
d to leverage photometric consistency across time when annotations are only avai
lable for a sparse subset of frames in a video. Our model is trained end-to-end
on color images to jointly reconstruct hands and objects in 3D by inferring thei
r poses. Given our estimated reconstructions, we differentiably render the optic
al flow between pairs of adjacent images and use it within the network to warp o
ne frame to another. We then apply a self-supervised photometric loss that relie
s on the visual consistency between nearby images. We achieve state-of-the-art r
esults on 3D hand-object reconstruction benchmarks and demonstrate that our appr
oach allows us to improve the pose estimation accuracy by leveraging information
 from neighboring frames in low-data regimes.
****

Scene Recomposition by Learning-Based ICP

Hamid Izadinia, Steven M. Seitz; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2020, pp. 930-939

By moving a depth sensor around a room, we compute a 3D CAD model of the environ
ment, capturing the room shape and contents such as chairs, desks, sofas, and ta
bles. Rather than reconstructing geometry, we match, place, and align each objec
t in the scene to thousands of CAD models of objects. In addition to the fully a
utomatic system, the key technical contribution is a novel approach for aligning
 CAD models to 3D scans, based on deep reinforcement learning. This approach, wh
ich we call Learning-based ICP, outperforms prior ICP methods in the literature,
 by learning the best points to match and conditioning on object viewpoint. LICP
 learns to align using only synthetic data and does not require ground truth ann
otation of object pose or keypoint pair matching in real scene scans. While LICP
 is trained on synthetic data and without 3D real scene annotations, it outperfo
rms both learned local deep feature matching and geometric based alignment metho
ds in real scenes. The proposed method is evaluated on real scenes datasets of S
ceneNN and ScanNet as well as synthetic scenes of SUNCG. High quality results ar
e demonstrated on a range of real world scenes, with robustness to clutter, view
point, and occlusion.
****

JL-DCF: Joint Learning and Densely-Cooperative Fusion Framework for RGB-D Salien
t Object Detection

Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3052-306
2

This paper proposes a novel joint learning and densely-cooperative fusion (JL-DC
F) architecture for RGB-D salient object detection. Existing models usually trea
t RGB and depth as independent information and design separate networks for feat
ure extraction from each. Such schemes can easily be constrained by a limited am
ount of training data or over-reliance on an elaborately-designed training proce
ss. In contrast, our JL-DCF learns from both RGB and depth inputs through a Siam
ese network. To this end, we propose two effective components: joint learning (J
L), and densely-cooperative fusion (DCF). The JL module provides robust saliency
 feature learning, while the latter is introduced for complementary feature disc

overy. Comprehensive experiments on four popular metrics show that the designed framework yields a robust RGB-D saliency detector with good generalization. As a result, JL-DCF significantly advances the top-1 D3Net model by an average of 1.9% (S-measure) across six challenging datasets, showing that the proposed framework offers a potential solution for real-world applications and could provide more insight into the cross-modality complementarity task. The code will be available at https://github.com/kerenfu/JLDCF/.

****

Inflated Episodic Memory With Region Self-Attention for Long-Tailed Visual Recognition

Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4344-4353

There have been increasing interests in modeling long-tailed data. Unlike artificially collected datasets, long-tailed data are naturally existed in the real-world and thus more realistic. To deal with the class imbalance problem, we introduce an Inflated Episodic Memory (IEM) for long-tailed visual recognition. First, our IEM augments the convolutional neural networks with categorical representative features for rapid learning on tail classes. In traditional few-shot learning, a single prototype is usually leveraged to represent a category. However, long-tailed data has higher intra-class variances. It could be challenging to learn a single prototype for one category. Thus, we introduce IEM to store the most discriminative feature for each category individually. Besides, the memory banks are updated independently, which further decreases the chance of learning skewed classifiers. Second, we introduce a novel region self-attention mechanism for multi-scale spatial feature map encoding. It is beneficial to incorporate more discriminative features to improve generalization on tail classes. We propose to encode local feature maps at multiple scales, and the spatial contextual information should be aggregated at the same time. Equipped with IEM and region self-attention, we achieve state-of-the-art performance on four standard long-tailed image recognition benchmarks. Besides, we validate the effectiveness of IEM on a long-tailed video recognition benchmark, i.e., YouTube-8M.

****

Learning to Select Base Classes for Few-Shot Classification

Linjun Zhou, Peng Cui, Xu Jia, Shiqiang Yang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4624-4633

Few-shot learning has attracted intensive research attention in recent years. Many methods have been proposed to generalize a model learned from provided base classes to novel classes, but no previous work studies how to select base classes, or even whether different base classes will result in different generalization performance of the learned model. In this paper, we utilize a simple yet effective measure, the Similarity Ratio, as an indicator for the generalization performance of a few-shot model. We then formulate the base class selection problem as a submodular optimization problem over Similarity Ratio. We further provide theoretical analysis on the optimization lower bound of different optimization methods, which could be used to identify the most appropriate algorithm for different experimental settings. The extensive experiments on ImageNet, Caltech256 and CUB-200-2011 demonstrate that our proposed method is effective in selecting a better base dataset.

****

Attention-Based Context Aware Reasoning for Situation Recognition

Thilini Cooray, Ngai-Man Cheung, Wei Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4736-4745

Situation Recognition (SR) is a fine-grained action recognition task where the model is expected to not only predict the salient action of the image, but also predict values of all associated semantic roles of the action. Predicting semantic roles is very challenging: a vast variety of possibilities can be the match for a semantic role. Existing work has focused on dependency modelling architectures to solve this issue. Inspired by the success achieved by query-based visual reasoning (e.g., Visual Question Answering), we propose to address semantic role

prediction as a query-based visual reasoning problem. However, existing query-based reasoning methods have not considered handling of inter-dependent queries which is a unique requirement of semantic role prediction in SR. Therefore, to the best of our knowledge, we propose the first set of methods to address inter-dependent queries in query-based visual reasoning. Extensive experiments demonstrate the effectiveness of our proposed method which achieves outstanding performance on Situation Recognition task. Furthermore, leveraging query inter-dependency, our methods improve upon a state-of-the-art method that answers queries separately. Our code: https://github.com/thilinicooray/context-aware-reasoning-for-sr
****

Towards Verifying Robustness of Neural Networks Against A Family of Semantic Perturbations
Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, Luca Daniel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 244-252
Verifying robustness of neural networks given a specified threat model is a fundamental yet challenging task. While current verification methods mainly focus on the l_p-norm threat model of the input instances, robustness verification against semantic adversarial attacks inducing large l_p-norm perturbations, such as color shifting and lighting adjustment, are beyond their capacity. To bridge this gap, we propose Semantify-NN, a model-agnostic and generic robustness verification approach against semantic perturbations for neural networks. By simply inserting our proposed semantic perturbation layers (SP-layers) to the input layer of any given model, Semantify-NN is model-agnostic, and any l_p-norm based verification tools can be used to verify the model robustness against semantic perturbations. We illustrate the principles of designing the SP-layers and provide examples including semantic perturbations to image classification in the space of hue, saturation, lightness, brightness, contrast and rotation, respectively. In addition, an efficient refinement technique is proposed to further significantly improve the semantic certificate. Experiments on various network architectures and different datasets demonstrate the superior verification performance of Semantify-NN over l_p-norm-based verification frameworks that naively convert semantic perturbation to l_p-norm. The results show that Semantify-NN can support robustness verification against a wide range of semantic perturbations.
****

Background Matting: The World Is Your Green Screen
Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M. Seitz, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2291-2300
We propose a method for creating a matte - the per-pixel foreground color and alpha - of a person by taking photos or videos in an everyday setting with a handheld camera. Most existing matting methods require a green screen background or a manually created trimap to produce a good matte. Automatic, trimap-free methods are appearing, but are not of comparable quality. In our trimap free approach, we ask the user to take an additional photo of the background without the subject at the time of capture. This step requires a small amount of foresight but is far less timeconsuming than creating a trimap. We train a deep network with an adversarial loss to predict the matte. We first train a matting network with a supervised loss on ground truth data with synthetic composites. To bridge the domain gap to real imagery with no labeling, we train another matting network guided by the first network and by a discriminator that judges the quality of composites. We demonstrate results on a wide variety of photos and videos and show significant improvement over the state of the art.
****

Learning to Simulate Dynamic Environments With GameGAN
Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1231-1240
Simulation is a crucial component of any robotic system. In order to simulate correctly, we need to write complex rules of the environment: how dynamic agents b

ehave, and how the actions of each of the agents affect the behavior of others. In this paper, we aim to learn a simulator by simply watching an agent interact with an environment. We focus on graphics games as a proxy of the real environment. We introduce GameGAN, a generative model that learns to visually imitate a desired game by ingesting screenplay and keyboard actions during training. Given a key pressed by the agent, GameGAN "renders" the next screen using a carefully designed generative adversarial network. Our approach offers key advantages over existing work: we design a memory module that builds an internal map of the environment, allowing for the agent to return to previously visited locations with high visual consistency. In addition, GameGAN is able to disentangle static and dynamic components within an image making the behavior of the model more interpretable, and relevant for downstream tasks that require explicit reasoning over dynamic elements. This enables many interesting applications such as swapping different components of the game to build new games that do not exist. We will release the code and trained model, enabling human players to play generated games and their variations with our GameGAN.

****

Active 3D Motion Visualization Based on Spatiotemporal Light-Ray Integration

Fumihiko Sakaue, Jun Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1980-1988

In this paper, we propose a method of visualizing 3D motion with zero latency. This method achieves motion visualization by projecting special high-frequency light patterns on moving objects without using any feedback mechanisms. For this objective, we focus on the time integration of light rays in the sensing system of observers. It is known that the visual system of human observers integrates light rays in a certain period. Similarly, the image sensor in a camera integrates light rays during the exposure time. Thus, our method embeds multiple images into a time-varying light field, such that the observer of the time-varying light field observes completely different images according to the dynamic motion of the scene. Based on this concept, we propose a method of generating special high-frequency patterns of projector lights. After projection onto target objects with projectors, the image observed on the target changes automatically depending on the motion of the objects and without any scene sensing and data analysis. In other words, we achieve motion visualization without the time delay incurred during sensing and computing.

****

Neural Networks Are More Productive Teachers Than Human Raters: Active Mixup for Data-Efficient Knowledge Distillation From a Blackbox Model

Dongdong Wang, Yandong Li, Liqiang Wang, Boqing Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1498-1507

We study how to train a student deep neural network for visual recognition by distilling knowledge from a blackbox teacher model in a data-efficient manner. Progress on this problem can significantly reduce the dependence on large-scale datasets for learning high-performing visual recognition models. There are two major challenges. One is that the number of queries into the teacher model should be minimized to save computational and/or financial costs. The other is that the number of images used for the knowledge distillation should be small; otherwise, it violates our expectation of reducing the dependence on large-scale datasets. To tackle these challenges, we propose an approach that blends mixup and active learning. The former effectively augments the few unlabeled images by a big pool of synthetic images sampled from the convex hull of the original images, and the latter actively chooses from the pool hard examples for the student neural network and query their labels from the teacher model. We validate our approach with extensive experiments.

****

Oops! Predicting Unintentional Action in Video

Dave Epstein, Boyuan Chen, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 919-929

From just a short glance at a video, we can often tell whether a person's action

is intentional or not. Can we train a model to recognize this? We introduce a dataset of in-the-wild videos of unintentional action, as well as a suite of tasks for recognizing, localizing, and anticipating its onset. We train a supervised neural network as a baseline and analyze its performance compared to human consistency on the tasks. We also investigate self-supervised representations that leverage natural signals in our dataset, and show the effectiveness of an approach that uses the intrinsic speed of video to perform competitively with highly-supervised pretraining. However, a significant gap between machine and human performance remains.
****

Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition
Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1112-1121
Skeleton-based human action recognition has attracted great interest thanks to the easy accessibility of the human skeleton data. Recently, there is a trend of using very deep feedforward neural networks to model the 3D coordinates of joints without considering the computational efficiency. In this paper, we propose a simple yet effective semantics-guided neural network (SGN) for skeleton-based action recognition. We explicitly introduce the high level semantics of joints (joint type and frame index) into the network to enhance the feature representation capability. In addition, we exploit the relationship of joints hierarchically through two modules, i.e., a joint-level module for modeling the correlations of joints in the same frame and a framelevel module for modeling the dependencies of frames by taking the joints in the same frame as a whole. A strong baseline is proposed to facilitate the study of this field. With an order of magnitude smaller model size than most previous works, SGN achieves the state-of-the-art performance on the NTU60, NTU120, and SYSU datasets.
****

Adversarial Vertex Mixup: Toward Better Adversarially Robust Generalization
Saehyung Lee, Hyungyu Lee, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 272-281
Adversarial examples cause neural networks to produce incorrect outputs with high confidence. Although adversarial training is one of the most effective forms of defense against adversarial examples, unfortunately, a large gap exists between test accuracy and training accuracy in adversarial training. In this paper, we identify Adversarial Feature Overfitting (AFO), which may cause poor adversarially robust generalization, and we show that adversarial training can overshoot the optimal point in terms of robust generalization, leading to AFO in our simple Gaussian model. Considering these theoretical results, we present soft labeling as a solution to the AFO problem. Furthermore, we propose Adversarial Vertex mixup (AVmixup), a soft-labeled data augmentation approach for improving adversarially robust generalization. We complement our theoretical analysis with experiments on CIFAR10, CIFAR100, SVHN, and Tiny ImageNet, and show that AVmixup significantly improves the robust generalization performance and that it reduces the trade-off between standard accuracy and adversarial robustness.
****

Learning to Autofocus
Charles Herrmann, Richard Strong Bowen, Neal Wadhwa, Rahul Garg, Qiurui He, Jonathan T. Barron, Ramin Zabih; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2230-2239
Autofocus is an important task for digital cameras, yet current approaches often exhibit poor performance. We propose a learning-based approach to this problem, and provide a realistic dataset of sufficient size for effective learning. Our dataset is labeled with per-pixel depths obtained from multi-view stereo, following [9]. Using this dataset, we apply modern deep classification models and an ordinal regression loss to obtain an efficient learning-based autofocus technique. We demonstrate that our approach provides a significant improvement compared with previous learned and non-learned methods: our model reduces the mean absolut

e error by a factor of 3.6 over the best comparable baseline algorithm. Our data set and code are publicly available.
****

Boosting the Transferability of Adversarial Samples via Attention
Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1161-1170
The widespread deployment of deep models necessitates the assessment of model vulnerability in practice, especially for safety- and security-sensitive domains such as autonomous driving and medical diagnosis. Transfer-based attacks against image classifiers thus elicit mounting interest, where attackers are required to craft adversarial images based on local proxy models without the feedback information from remote target ones. However, under such a challenging but practical setup, the synthesized adversarial samples often achieve limited success due to overfitting to the local model employed. In this work, we propose a novel mechanism to alleviate the overfitting issue. It computes model attention over extracted features to regularize the search of adversarial examples, which prioritizes the corruption of critical features that are likely to be adopted by diverse architectures. Consequently, it can promote the transferability of resultant adversarial instances. Extensive experiments on ImageNet classifiers confirm the effectiveness of our strategy and its superiority to state-of-the-art benchmarks in both white-box and black-box settings.
****

Single Image Reflection Removal Through Cascaded Refinement
Chao Li, Yixiao Yang, Kun He, Stephen Lin, John E. Hopcroft; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3565-3574
We address the problem of removing undesirable reflections from a single image captured through a glass surface, which is an ill-posed, challenging but practically important problem for photo enhancement. Inspired by iterative structure reduction for hidden community detection in social networks, we propose an Iterative Boost Convolutional LSTM Network (IBCLN) that enables cascaded prediction for reflection removal. IBCLN is a cascaded network that iteratively refines the estimates of transmission and reflection layers in a manner that they can boost the prediction quality to each other, and information across steps of the cascade is transferred using an LSTM. The intuition is that the transmission is the strong, dominant structure while the reflection is the weak, hidden structure. They are complementary to each other in a single image and thus a better estimate and reduction on one side from the original image leads to a more accurate estimate on the other side. To facilitate training over multiple cascade steps, we employ LSTM to address the vanishing gradient problem, and propose residual reconstruction loss as further training guidance. Besides, we create a dataset of real-world images with reflection and ground-truth transmission layers to mitigate the problem of insufficient data. Comprehensive experiments demonstrate that the proposed method can effectively remove reflections in real and synthetic images compared with state-of-the-art reflection removal methods.
****

Distilled Semantics for Comprehensive Scene Understanding from Videos
Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, Stefano Mattoccia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4654-4665
Whole understanding of the surroundings is paramount to autonomous systems. Recent works have shown that deep neural networks can learn geometry (depth) and motion (optical flow) from a monocular video without any explicit supervision from ground truth annotations, particularly hard to source for these two tasks. In this paper, we take an additional step toward holistic scene understanding with monocular cameras by learning depth and motion alongside with semantics, with supervision for the latter provided by a pre-trained network distilling proxy ground truth images. We address the three tasks jointly by a) a novel training protocol based on knowledge distillation and self-supervision and b) a compact network

architecture which enables efficient scene understanding on both power hungry GPUs and low-power embedded platforms. We thoroughly assess the performance of our framework and show that it yields state-of-the-art results for monocular depth estimation, optical flow and motion segmentation.
****

Multi-Dimensional Pruning: A Unified Framework for Model Compression
Jinyang Guo, Wanli Ouyang, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1508-1517
In this work, we propose a unified model compression framework called Multi-Dimensional Pruning (MDP) to simultaneously compress the convolutional neural networks (CNNs) on multiple dimensions. In contrast to the existing model compression methods that only aim to reduce the redundancy along either the spatial/spatial-temporal dimension (e.g., spatial dimension for 2D CNNs, spatial and temporal dimensions for 3D CNNs) or the channel dimension, our newly proposed approach can simultaneously reduce the spatial/spatial-temporal and the channel redundancies for CNNs. Specifically, in order to reduce the redundancy along the spatial/spatial-temporal dimension, we downsample the input tensor of a convolutional layer, in which the scaling factor for the downsampling operation is adaptively selected by our approach. After the convolution operation, the output tensor is upsampled to the original size to ensure the unchanged input size for the subsequent CNN layers. To reduce the channel-wise redundancy, we introduce a gate for each channel of the output tensor as its importance score, in which the gate value is automatically learned. The channels with small importance scores will be removed after the model compression process. Our comprehensive experiments on four benchmark datasets demonstrate that our MDP framework outperforms the existing methods when pruning both 2D CNNs and 3D CNNs.
****

Varicolored Image De-Hazing
Akshay Dudhane, Kuldeep M. Biradar, Prashant W. Patil, Praful Hambarde, Subrahmanyam Murala; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4564-4573
The quality of images captured in bad weather is often affected by chromatic casts and low visibility due to the presence of atmospheric particles. Restoration of the color balance is often ignored in most of the existing image de-hazing methods. In this paper, we propose a varicolored end-to-end image de-hazing network which restores the color balance in a given varicolored hazy image and recovers the haze-free image. The proposed network comprises of 1) Haze color correction (HCC) module and 2) Visibility improvement (VI) module. The proposed HCC module provides required attention to each color channel and generates a color balanced hazy image. While the proposed VI module processes the color balanced hazy image through novel inception attention block to recover the haze-free image. We also propose a novel approach to generate a large-scale varicolored synthetic hazy image database. An ablation study has been carried out to demonstrate the effect of different factors on the performance of the proposed network for image de-hazing. Three benchmark synthetic datasets have been used for quantitative analysis of the proposed network. Visual results on a set of real-world hazy images captured in different weather conditions demonstrate the effectiveness of the proposed approach for varicolored image de-hazing.
****

Visually Imbalanced Stereo Matching
Yicun Liu, Jimmy Ren, Jiawei Zhang, Jianbo Liu, Mude Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2029-2038
Understanding of human vision system (HVS) has inspired many computer vision algorithms. Stereo matching, which borrows the idea from human stereopsis, has been extensively studied in the existing literature. However, scant attention has been drawn on a typical scenario where binocular inputs are qualitatively different (e.g., high-res master camera and low-res slave camera in a dual-lens module). Recent advances in human optometry reveal the capability of the human visual system to maintain coarse stereopsis under such visually imbalanced conditions. Bi

onically aroused, it is natural to question that: do stereo machines share the same capability? In this paper, we carry out a systematic comparison to investigate the effect of various imbalanced conditions on current popular stereo matching algorithms. We show that resembling the human visual system, those algorithms can handle limited degrees of monocular downgrading but also prone to collapses beyond a certain threshold. To avoid such collapse, we propose a solution to recover the stereopsis by a joint guided-view-restoration and stereo-reconstruction framework. We show the superiority of our framework on KITTI dataset and its extension on real-world applications.
****

Defending Against Model Stealing Attacks With Adaptive Misinformation
Sanjay Kariyappa, Moinuddin K. Qureshi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 770-778
Deep Neural Networks (DNNs) are susceptible to model stealing attacks, which allows a data-limited adversary with no knowledge of the training dataset to clone the functionality of a target model, just by using black-box query access. Such attacks are typically carried out by querying the target model using inputs that are synthetically generated or sampled from a surrogate dataset to construct a labeled dataset. The adversary can use this labeled dataset to train a clone model, which achieves a classification accuracy comparable to that of the target model. We propose "Adaptive Misinformation" to defend against such model stealing attacks. We identify that all existing model stealing attacks invariably query the target model with Out-Of-Distribution (OOD) inputs. By selectively sending incorrect predictions for OOD queries, our defense substantially degrades the accuracy of the attacker's clone model (by up to 40%), while minimally impacting the accuracy (<0.5%) for benign users. Compared to existing defenses, our defense has a significantly better security vs accuracy trade-off and incurs minimal computational overhead.
****

Semi-Supervised Learning for Few-Shot Image-to-Image Translation
Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4453-4462
In the last few years, unpaired image-to-image translation has witnessed Remarkable progress. Although the latest methods are able to generate realistic images, they crucially rely on a large number of labeled images. Recently, some methods have tackled the challenging setting of few-shot image-to-image ranslation, reducing the labeled data requirements for the target domain during inference. In this work, we go one step further and reduce the amount of required labeled data also from the source domain during training. To do so, we propose applying semi-supervised learning via a noise-tolerant pseudo-labeling procedure. We also apply a cycle consistency constraint to further exploit the information from unlabeled images, either from the same dataset or external. Additionally, we propose several structural modifications to facilitate the image translation task under these circumstances. Our semi-supervised method for few-shot image translation, called SEMIT, achieves excellent results on four different datasets using as little as 10% of the source labels, and matches the performance of the main fully-supervised competitor using only 20% labeled data. Our code and models are made public at: https://github.com/yaxingwang/SEMIT.
****

Single Image Optical Flow Estimation With an Event Camera
Liyuan Pan, Miaomiao Liu, Richard Hartley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1672-1681
Event cameras are bio-inspired sensors that asynchronously report intensity changes in microsecond resolution. DAVIS can capture high dynamics of a scene and simultaneously output high temporal resolution events and low frame-rate intensity images. In this paper, we propose a single image (potentially blurred) and events based optical flow estimation approach. First, we demonstrate how events can be used to improve flow estimates. To this end, we encode the relation between flow and events effectively by presenting an event-based photometric consistency

formulation. Then, we consider the special case of image blur caused by high dyn amics in the visual environments and show that including the blur formation in o ur model further constrains flow estimation. This is in sharp contrast to existi ng works that ignore the blurred images while our formulation can naturally hand le either blurred or sharp images to achieve accurate flow estimation. Finally, we reduce flow estimation, as well as image deblurring, to an alternative optimi zation problem of an objective function using the primal-dual algorithm. Experim ental results on both synthetic and real data (with blurred and non-blurred imag es) show the superiority of our model in comparison to state-of-the-art approach es.
****

Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention
Dat Huynh, Ehsan Elhamifar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4483-4493
We address the problem of fine-grained generalized zero-shot recognition of visu ally similar classes without training images for some classes. We propose a dens e attribute-based attention mechanism that for each attribute focuses on the mos t relevant image regions, obtaining attribute-based features. Instead of alignin g a global feature vector of an image with its associated class semantic vector, we propose an attribute embedding technique that aligns each attribute-based fe ature with its attribute semantic vector. Hence, we compute a vector of attribut e scores, for the presence of each attribute in an image, whose similarity with the true class semantic vector is maximized. Moreover, we adjust each attribute score using an attention mechanism over attributes to better capture the discrim inative power of different attributes. To tackle the challenge of bias towards s een classes during testing, we propose a new self-calibration loss that adjusts the probability of unseen classes to account for the training bias. We conduct e xperiments on three popular datasets of CUB, SUN and AWA2 as well as the large-s cale DeepFashion dataset, showing that our model significantly improves the stat e of the art.
****

Solving Jigsaw Puzzles With Eroded Boundaries
Dov Bridger, Dov Danon, Ayellet Tal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3526-3535
Jigsaw puzzle solving is an intriguing problem which has been explored in comput er vision for decades. This paper focuses on a specific variant of the problem-- solving puzzles with eroded boundaries. Such erosion makes the problem extremely difficult, since most existing solvers utilize solely the information at the bo undaries. Nevertheless, this variant is important since erosion and missing data often occur at the boundaries. The key idea of our proposed approach is to inpa int the eroded boundaries between puzzle pieces and later leverage the quality o f the inpainted area to classify a pair of pieces as "neighbors or not". An inte resting feature of our architecture is that the same GAN discriminator is used f or both inpainting and classification; training of the second task is simply a c ontinuation of the training of the first, beginning from the point it left off. We show that our approach outperforms other SOTA methods.
****

Learning Rank-1 Diffractive Optics for Single-Shot High Dynamic Range Imaging
Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, Felix Heide; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1386-1396
High-dynamic range (HDR) imaging is an essential imaging modality for a wide ran ge of applications in uncontrolled environments, including autonomous driving, r obotics, and mobile phone cameras. However, existing HDR techniques in commodity devices struggle with dynamic scenes due to multi-shot acquisition and post-pro cessing time, e.g. mobile phone burst photography, making such approaches unsuit able for real-time applications. In this work, we propose a method for snapshot HDR imaging by learning an optical HDR encoding in a single image which maps sat urated highlights into neighboring unsaturated areas using a diffractive optical element (DOE). We propose a novel rank-1 parameterization of the proposed DOE w

hich avoids vast trainable parameters and keeps high frequencies' encoding compa
red with conventional end-to-end design methods. We further propose a reconstruc
tion network tailored to this rank-1 parametrization for recovery of clipped inf
ormation from the encoded measurements. The proposed end-to-end framework is val
idated through simulation and real-world experiments and improves the PSNR by mo
re than 7 dB over state-of-the-art end-to-end designs.
****

Universal Source-Free Domain Adaptation
Jogendra Nath Kundu, Naveen Venkat, Rahul M V, R. Venkatesh Babu; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
020, pp. 4544-4553
There is a strong incentive to develop versatile learning techniques that can tr
ansfer the knowledge of class-separability from a labeled source domain to an un
labeled target domain in the presence of a domain-shift. Existing domain adaptat
ion (DA) approaches are not equipped for practical DA scenarios as a result of t
heir reliance on the knowledge of source-target label-set relationship (e.g. Clo
sed-set, Open-set or Partial DA). Furthermore, almost all prior unsupervised DA
works require coexistence of source and target samples even during deployment, m
aking them unsuitable for real-time adaptation. Devoid of such impractical assum
ptions, we propose a novel two-stage learning process. 1) In the Procurement sta
ge, we aim to equip the model for future source-free deployment, assuming no pri
or knowledge of the upcoming category-gap and domain-shift. To achieve this, we
enhance the model's ability to reject out-of-source distribution samples by leve
raging the available source data, in a novel generative classifier framework. 2)
 In the Deployment stage, the goal is to design a unified adaptation algorithm c
apable of operating across a wide range of category-gaps, with no access to the
previously seen source samples. To this end, in contrast to the usage of complex
 adversarial training regimes, we define a simple yet effective source-free adap
tation objective by utilizing a novel instance-level weighting mechanism, named
as Source Similarity Metric (SSM). A thorough evaluation shows the practical usa
bility of the proposed learning framework with superior DA performance even over
 state-of-the-art source-dependent approaches.
****

Meta-Transfer Learning for Zero-Shot Super-Resolution
Jae Woong Soh, Sunwoo Cho, Nam Ik Cho; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3516-3525
Convolutional neural networks (CNNs) have shown dramatic improvements in single
image super-resolution (SISR) by using large-scale external samples. Despite the
ir remarkable performance based on the external dataset, they cannot exploit int
ernal information within a specific image. Another problem is that they are appl
icable only to the specific condition of data that they are supervised. For inst
ance, the low-resolution (LR) image should be a "bicubic" downsampled noise-free
 image from a high-resolution (HR) one. To address both issues, zero-shot super-
resolution (ZSSR) has been proposed for flexible internal learning. However, the
y require thousands of gradient updates, i.e., long inference time. In this pape
r, we present Meta-Transfer Learning for Zero-Shot Super-Resolution (MZSR), whic
h leverages ZSSR. Precisely, it is based on finding a generic initial parameter
that is suitable for internal learning. Thus, we can exploit both external and i
nternal information, where one single gradient update can yield quite considerab
le results. With our method, the network can quickly adapt to a given image cond
ition. In this respect, our method can be applied to a large spectrum of image c
onditions within a fast adaptation process.
****

A Model-Driven Deep Neural Network for Single Image Rain Removal
Hong Wang, Qi Xie, Qian Zhao, Deyu Meng; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3103-3112
Deep learning (DL) methods have achieved state-of-the-art performance in the tas
k of single image rain removal. Most of current DL architectures, however, are s
till lack of sufficient interpretability and not fully integrated with physical
structures inside general rain streaks. To this issue, in this paper, we propose

a model-driven deep neural network for the task, with fully interpretable network structures. Specifically, based on the convolutional dictionary learning mechanism for representing rain, we propose a novel single image deraining model and utilize the proximal gradient descent technique to design an iterative algorithm only containing simple operators for solving the model. Such a simple implementation scheme facilitates us to unfold it into a new deep network architecture, called rain convolutional dictionary network (RCDNet), with almost every network module one-to-one corresponding to each operation involved in the algorithm. By end-to-end training the proposed RCDNet, all the rain kernels and proximal operators can be automatically extracted, faithfully characterizing the features of both rain and clean background layers, and thus naturally lead to its better deraining performance, especially in real scenarios. Comprehensive experiments substantiate the superiority of the proposed network, especially its well generality to diverse testing scenarios and good interpretability for all its modules, as compared with state-of-the-arts both visually and quantitatively.
****

Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution

Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, Jianhua Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4837-4846

Multiple instance learning (MIL) is a typical weakly-supervised learning method where the label is associated with a bag of instances instead of a single instance. Despite extensive research over past years, effectively deploying MIL remains an open and challenging problem, especially when the commonly assumed standard multiple instance (SMI) assumption is not satisfied. In this paper, we propose a multiple instance learning method based on deep graph convolutional network and feature selection (FS-GCN-MIL) for histopathological image classification. The proposed method consists of three components, including instance-level feature extraction, instance-level feature selection, and bag-level classification. We develop a self-supervised learning mechanism to train the feature extractor based on a combination model of variational autoencoder and generative adversarial network (VAE-GAN). Additionally, we propose a novel instance-level feature selection method to select the discriminative instance features. Furthermore, we employ a graph convolutional network (GCN) for learning the bag-level representation and then performing the classification. We apply the proposed method in the prediction of lymph node metastasis using histopathological images of colorectal cancer. Experimental results demonstrate that the proposed method achieves superior performance compared to the state-of-the-art methods.
****

APQ: Joint Search for Network Architecture, Pruning and Quantization Policy

Tianzhe Wang, Kuan Wang, Han Cai, Ji Lin, Zhijian Liu, Hanrui Wang, Yujun Lin, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2078-2087

We present APQ, a novel design methodology for efficient deep learning deployment. Unlike previous methods that separately optimize the neural network architecture, pruning policy, and quantization policy, we design to optimize them in a joint manner. To deal with the larger design space it brings, we devise to train a quantization-aware accuracy predictor that is fed to the evolutionary search to select the best fit. Since directly training such a predictor requires time-consuming quantization data collection, we propose to use predictor-transfer technique to get the quantization-aware predictor: we first generate a large dataset of pairs by sampling a pretrained unified supernet and doing direct evaluation; then we use these data to train an accuracy predictor without quantization, further transferring its weights to train the quantization-aware predictor, which largely reduces the quantization data collection time. Extensive experiments on ImageNet show the benefits of this joint design methodology: the model searched by our method maintains the same level accuracy as ResNet34 8-bit model while saving 8x BitOps; we obtain the same level accuracy as MobileNetV2+HAQ while achievi

ng 2x/1.3x latency/energy saving; the marginal search cost of joint optimization for a new deployment scenario outperforms separate optimizations using ProxylessNAS+AMC+HAQ by 2.3% accuracy while reducing 600x GPU hours and CO2 emission.
****

SSRNet: Scalable 3D Surface Reconstruction Network
Zhenxing Mi, Yiming Luo, Wenbing Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 970-979
Existing learning-based surface reconstruction methods from point clouds are still facing challenges in terms of scalability and preservation of details on large-scale point clouds. In this paper, we propose the SSRNet, a novel scalable learning-based method for surface reconstruction. The proposed SSRNet constructs local geometry-aware features for octree vertices and designs a scalable reconstruction pipeline, which not only greatly enhances the predication accuracy of the relative position between the vertices and the implicit surface facilitating the surface reconstruction quality, but also allows dividing the point cloud and octree vertices and processing different parts in parallel for superior scalability on large-scale point clouds with millions of points. Moreover, SSRNet demonstrates outstanding generalization capability and only needs several surface data for training, much less than other learning-based reconstruction methods, which can effectively avoid overfitting. The trained model of SSRNet on one dataset can be directly used on other datasets with superior performance. Finally, the time consumption with SSRNet on a large-scale point cloud is acceptable and competitive. To our knowledge, the proposed SSRNet is the first to really bring a convincing solution to the scalability issue of the learning-based surface reconstruction methods, and is an important step to make learning-based methods competitive with respect to geometry processing methods on real-world and challenging data. Experiments show that our method achieves a breakthrough in scalability and quality compared with state-of-the-art learning-based methods.
****

Semantic Correspondence as an Optimal Transport Problem
Yanbin Liu, Linchao Zhu, Makoto Yamada, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4463-4472
Establishing dense correspondences across semantically similar images is a challenging task. Due to the large intra-class variation and background clutter, two common issues occur in current approaches. First, many pixels in a source image are assigned to one target pixel, i.e., many to one matching. Second, some object pixels are assigned to the background pixels, i.e., background matching. We solve the first issue by global feature matching, which maximizes the total matching correlations between images to obtain a global optimal matching matrix. The row sum and column sum constraints are enforced on the matching matrix to induce a balanced solution, thus suppressing the many to one matching. We solve the second issue by applying a staircase function on the class activation maps to re-weight the importance of pixels into four levels from foreground to background. The whole procedure is combined into a unified optimal transport algorithm by converting the maximization problem to the optimal transport formulation and incorporating the staircase weights into optimal transport algorithm to act as empirical distributions. The proposed algorithm achieves state-of-the-art performance on four benchmark datasets. Notably, a 26% relative improvement is achieved on the large-scale SPair-71k dataset.
****

Improving Confidence Estimates for Unfamiliar Examples
Zhizhong Li, Derek Hoiem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2686-2695
Intuitively, unfamiliarity should lead to lack of confidence. In reality, current algorithms often make highly confident yet wrong predictions when faced with relevant but unfamiliar examples. A classifier we trained to recognize gender is 12 times more likely to be wrong with a 99% confident prediction if presented with a subject from a different age group than those seen during training. In this paper, we compare and evaluate several methods to improve confidence estimates

for unfamiliar and familiar samples. We propose a testing methodology of splitti
ng unfamiliar and familiar samples by attribute (age, breed, subcategory) or sam
pling (similar datasets collected by different people at different times). We ev
aluate methods including confidence calibration, ensembles, distillation, and a
Bayesian model and use several metrics to analyze label, likelihood, and calibra
tion error. While all methods reduce over-confident errors, the ensemble of cali
brated models performs best overall, and T-scaling performs best among the appro
aches with fastest inference.
****

Learning Generative Models of Shape Handles
Matheus Gadelha, Giorgio Gori, Duygu Ceylan, Radomir Mech, Nathan Carr, Tam
y Boubekeur, Rui Wang, Subhransu Maji; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 402-411
We present a generative model to synthesize 3D shapes as sets of handles -- ligh
tweight proxies that approximate the original 3D shape -- for applications in in
teractive editing, shape parsing, and building compact 3D representations. Our m
odel can generate handle sets with varying cardinality and different types of ha
ndles. Key to our approach is a deep architecture that predicts both the paramet
ers and existence of shape handles and a novel similarity measure that can easil
y accommodate different types of handles, such as cuboids or sphere-meshes. We l
everage the recent advances in semantic 3D annotation as well as automatic shape
summarization techniques to supervise our approach. We show that the resulting
shape representations are not only intuitive, but achieve superior quality than
previous state-of-the-art. Finally, we demonstrate how our method can be used in
applications such as interactive shape editing and completion, leveraging the l
atent space learned by our model to guide these tasks.
****

Toward a Universal Model for Shape From Texture
Dor Verbin, Todd Zickler; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2020, pp. 422-430
We consider the shape from texture problem, where the input is a single image of
a curved, textured surface, and the texture and shape are both a priori unknown
. We formulate this task as a three-player game between a shape process, a textu
re process, and a discriminator. The discriminator adapts a set of non-linear fi
lters to try to distinguish image patches created by the texture process from th
ose created by the shape process, while the shape and texture processes try to c
reate image patches that are indistinguishable from those of the other. An equil
ibrium of this game yields two things: an estimate of the 2.5D surface from the
shape process, and a stochastic texture synthesis model from the texture process
. Experiments show that this approach is robust to common non-idealities such as
shading, gloss, and clutter. We also find that it succeeds for a wide variety o
f texture types, including both periodic textures and those composed of isolated
textons, which have previously required distinct and specialized processing.
****

Stochastic Sparse Subspace Clustering
Ying Chen, Chun-Guang Li, Chong You; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4155-4164
State-of-the-art subspace clustering methods are based on self-expressive model,
which represents each data point as a linear combination of other data points.
By enforcing such representation to be sparse, sparse subspace clustering is gua
ranteed to produce a subspace-preserving data affinity where two points are conn
ected only if they are from the same subspace. On the other hand, however, data
points from the same subspace may not be well-connected, leading to the issue of
over-segmentation. We introduce dropout to address the issue of over-segmentati
on, which is based on randomly dropping out data points in self-expressive model
. In particular, we show that dropout is equivalent to adding a squared l_2 norm
regularization on the representation coefficients, therefore induces denser sol
utions. Then, we reformulate the optimization problem as a consensus problem ove
r a set of small-scale subproblems. This leads to a scalable and flexible sparse
subspace clustering approach, termed Stochastic Sparse Subspace Clustering, whi

ch can effectively handle large scale datasets. Extensive experiments on synthet ic data and real world datasets validate the efficiency and effectiveness of our proposal.
****

Learn2Perturb: An End-to-End Feature Perturbation Learning to Improve Adversaria l Robustness
Ahmadreza Jeddi, Mohammad Javad Shafiee, Michelle Karg, Christian Scharfenber ger, Alexander Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1241-1250
While deep neural networks have been achieving state-of-the-art performance acro ss a wide variety of applications, their vulnerability to adversarial attacks li mits their widespread deployment for safety-critical applications. Alongside oth er adversarial defense approaches being investigated, there has been a very rece nt interest in improving adversarial robustness in deep neural networks through the introduction of perturbations during the training process. However, such met hods leverage fixed, pre-defined perturbations and require significant hyper-par ameter tuning that makes them very difficult to leverage in a general fashion. I n this study, we introduce Learn2Perturb, an end-to-end feature perturbation lea rning approach for improving the adversarial robustness of deep neural networks. More specifically, we introduce novel perturbation-injection modules that are i ncorporated at each layer to perturb the feature space and increase uncertainty in the network. This feature perturbation is performed at both the training and the inference stages. Furthermore, inspired by the Expectation-Maximization, an alternating back-propagation training algorithm is introduced to train the netwo rk and noise parameters consecutively. Experimental results on CIFAR-10 and CIFA R-100 datasets show that the proposed Learn2Perturb method can result in deep ne ural networks which are 4-7% more robust on l_inf FGSM and PDG adversarial attac ks and significantly outperforms the state-of-the-art against l_2 C&W attack and a wide range of well-known black-box attacks.
****

Syn2Real Transfer Learning for Image Deraining Using Gaussian Processes
Rajeev Yasarla, Vishwanath A. Sindagi, Vishal M. Patel; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 27 26-2736
Recent CNN-based methods for image deraining have achieved excellent performance in terms of reconstruction error as well as visual quality. However, these meth ods are limited in the sense that they can be trained only on fully labeled data . Due to various challenges in obtaining real world fully-labeled image derainin g datasets, existing methods are trained only on synthetically generated data an d hence, generalize poorly to real-world images. The use of real-world data in t raining image deraining networks is relatively less explored in the literature. We propose a Gaussian Process-based semi-supervised learning framework which ena bles the network in learning to derain using synthetic dataset while generalizin g better using unlabeled real-world images. Through extensive experiments and ab lations on several challenging datasets (such as Rain800, Rain100H and DDN-SIRR) , we show that the proposed method, when trained on limited labeled data, achiev es on-par performance with fully-labeled training. Additionally, we demonstrate that using unlabeled real-world images in the proposed GP-based framework result s in superior performance as compared to existing methods.
****

On Isometry Robustness of Deep 3D Point Cloud Models Under Adversarial Attacks
Yue Zhao, Yuwei Wu, Caihua Chen, Andrew Lim; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1201-1210
While deep learning in 3D domain has achieved revolutionary performance in many tasks, the robustness of these models has not been sufficiently studied or explo red. Regarding the 3D adversarial samples, most existing works focus on manipula tion of local points, which may fail to invoke the global geometry properties, l ike robustness under linear projection that preserves the Euclidean distance, i. e., isometry. In this work, we show that existing state-of-the-art deep 3D model s are extremely vulnerable to isometry transformations. Armed with the Thompson

Sampling, we develop a black-box attack with success rate over 95% on ModelNet40 data set. Incorporating with the Restricted Isometry Property, we propose a novel framework of white-box attack on top of spectral norm based perturbation. In contrast to previous works, our adversarial samples are experimentally shown to be strongly transferable. Evaluated on a sequence of prevailing 3D models, our white-box attack achieves success rates from 98.88% to 100%. It maintains a successful attack rate over 95% even within an imperceptible rotation range [+-2.81*].

****

Self-Supervised Viewpoint Learning From Image Collections

Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3971-3981

Training deep neural networks to estimate the viewpoint of objects requires large labeled training datasets. However, manually labeling viewpoints is notoriously hard, error-prone, and time-consuming. On the other hand, it is relatively easy to mine many unlabeled images of an object category from the internet, e.g., of cars or faces. We seek to answer the research question of whether such unlabeled collections of in-the-wild images can be successfully utilized to train viewpoint estimation networks for general object categories purely via self-supervision. Self-supervision here refers to the fact that the only true supervisory signal that the network has is the input image itself. We propose a novel learning framework which incorporates an analysis-by-synthesis paradigm to reconstruct images in a viewpoint aware manner with a generative network, along with symmetry and adversarial constraints to successfully supervise our viewpoint estimation network. We show that our approach performs competitively to fully-supervised approaches for several object categories like human faces, cars, buses, and trains. Our work opens up further research in self-supervised viewpoint learning and serves as a robust baseline for it. We open-source our code at https://github.com/NVlabs/SSV.

****

On the Uncertainty of Self-Supervised Monocular Depth Estimation

Matteo Poggi, Filippo Aleotti, Fabio Tosi, Stefano Mattoccia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3227-3237

Self-supervised paradigms for monocular depth estimation are very appealing since they do not require ground truth annotations at all. Despite the astonishing results yielded by such methodologies, learning to reason about the uncertainty of the estimated depth maps is of paramount importance for practical applications, yet uncharted in the literature. Purposely, we explore for the first time how to estimate the uncertainty for this task and how this affects depth accuracy, proposing a novel peculiar technique specifically designed for self-supervised approaches. On the standard KITTI dataset, we exhaustively assess the performance of each method with different self-supervised paradigms. Such evaluation highlights that our proposal i) always improves depth accuracy significantly and ii) yields state-of-the-art results concerning uncertainty estimation when training on sequences and competitive results uniquely deploying stereo pairs.

****

StegaStamp: Invisible Hyperlinks in Physical Photographs

Matthew Tancik, Ben Mildenhall, Ren Ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2117-2126

Printed and digitally displayed photos have the ability to hide imperceptible digital data that can be accessed through internet-connected imaging systems. Another way to think about this is physical photographs that have unique QR codes invisibly embedded within them. This paper presents an architecture, algorithms, and a prototype implementation addressing this vision. Our key technical contribution is StegaStamp, a learned steganographic algorithm to enable robust encoding and decoding of arbitrary hyperlink bitstrings into photos in a manner that approaches perceptual invisibility. StegaStamp comprises a deep neural network that learns an encoding/decoding algorithm robust to image perturbations approximati

ng the space of distortions resulting from real printing and photography. We dem
onstrates real-time decoding of hyperlinks in photos from in-the-wild videos tha
t contain variation in lighting, shadows, perspective, occlusion and viewing dis
tance. Our prototype system robustly retrieves 56 bit hyperlinks after error cor
rection -- sufficient to embed a unique code within every photo on the internet.
****

Anisotropic Convolutional Networks for 3D Semantic Scene Completion
Jie Li, Kai Han, Peng Wang, Yu Liu, Xia Yuan; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3351-3359
As a voxel-wise labeling task, semantic scene completion (SSC) tries to simultan
eously infer the occupancy and semantic labels for a scene from a single depth a
nd/or RGB image. The key challenge for SSC is how to effectively take advantage
of the 3D context to model various objects or stuffs with severe variations in s
hapes, layouts, and visibility. To handle such variations, we propose a novel mo
dule called anisotropic convolution, which properties with flexibility and power
impossible for the competing methods such as standard 3D convolution and some o
f its variations. In contrast to the standard 3D convolution that is limited to
a fixed 3D receptive field, our module is capable of modeling the dimensional an
isotropy voxel-wisely. The basic idea is to enable anisotropic 3D receptive fiel
d by decomposing a 3D convolution into three consecutive 1D convolutions, and th
e kernel size for each such 1D convolution is adaptively determined on the fly.
By stacking multiple such anisotropic convolution modules, the voxel-wise modeli
ng capability can be further enhanced while maintaining a controllable amount of
model parameters. Extensive experiments on two SSC benchmarks, NYU-Depth-v2 and
NYUCAD, show the superior performance of the proposed method.
****

Learning to Have an Ear for Face Super-Resolution
Givi Meishvili, Simon Jenni, Paolo Favaro; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1364-1374
We propose a novel method to use both audio and a low-resolution image to perfor
m extreme face super-resolution (a 16x increase of the input size). When the res
olution of the input image is very low (e.g., 8x8 pixels), the loss of informati
on is so dire that important details of the original identity have been lost and
audio can aid the recovery of a plausible high-resolution image. In fact, audio
carries information about facial attributes, such as gender and age. To combine
the aural and visual modalities, we propose a method to first build the latent
representations of a face from the lone audio track and then from the lone low-r
esolution image. We then train a network to fuse these two representations. We s
how experimentally that audio can assist in recovering attributes such as the ge
nder, the age and the identity, and thus improve the correctness of the high-res
olution image reconstruction process. Our procedure does not make use of human a
nnotation and thus can be easily trained with existing video datasets. Moreover,
we show that our model builds a factorized representation of images and audio a
s it allows one to mix low-resolution images and audio from different videos and
to generate realistic faces with semantically meaningful combinations.
****

Cascaded Refinement Network for Point Cloud Completion
Xiaogang Wang, Marcelo H. Ang Jr., Gim Hee Lee; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 790-799
Point clouds are often sparse and incomplete. Existing shape completion methods
are incapable of generating details of objects or learning the complex point dis
tributions. To this end, we propose a cascaded refinement network together with
a coarse-to-fine strategy to synthesize the detailed object shapes. Considering
the local details of partial input with the global shape information together, w
e can preserve the existing details in the incomplete point set and generate the
missing parts with high fidelity. We also design a patch discriminator that gua
rantees every local area has the same pattern with the ground truth to learn the
complicated point distribution. Quantitative and qualitative experiments on dif
ferent datasets show that our method achieves superior results compared to exist
ing state-of-the-art approaches on the 3D point cloud completion task. Our sourc

e code is available at https://github.com/xiaogangw/cascaded-point-completion.git.
****
DOA-GAN: Dual-Order Attentive Generative Adversarial Network for Image Copy-Move Forgery Detection and Localization
Ashraful Islam, Chengjiang Long, Arslan Basharat, Anthony Hoogs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4676-4685
Images can be manipulated for nefarious purposes to hide content or to duplicate certain objects through copy-move operations. Discovering a well-crafted copy-move forgery in images can be very challenging for both humans and machines; for example, an object on a uniform background can be replaced by an image patch of the same background. In this paper, we propose a Generative Adversarial Network with a dual-order attention model to detect and localize copy-move forgeries. In the generator, the first-order attention is designed to capture copy-move location information, and the second-order attention exploits more discriminative features for the patch co-occurrence. Both attention maps are extracted from the affinity matrix and are used to fuse location-aware and co-occurrence features for the final detection and localization branches of the network. The discriminator network is designed to further ensure more accurate localization results. To the best of our knowledge, we are the first to propose such a network architecture with the 1st-order attention mechanism from the affinity matrix. We have performed extensive experimental validation and our state-of-the-art results strongly demonstrate the efficacy of the proposed approach.
****
Rotation Equivariant Graph Convolutional Network for Spherical Image Classification
Qin Yang, Chenglin Li, Wenrui Dai, Junni Zou, Guo-Jun Qi, Hongkai Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4303-4312
Convolutional neural networks (CNNs) designed for low-dimensional regular grids will unfortunately lead to non-optimal solutions for analyzing spherical images, due to their different geometrical properties from planar images. In this paper, we generalize the grid-based CNNs to a non-Euclidean space by taking into account the geometry of spherical surfaces and propose a Spherical Graph Convolutional Network (SGCN) to encode rotation equivariant representations. Specifically, we propose a spherical graph construction criterion showing that a graph needs to be regular by evenly covering the spherical surfaces in order to design a rotation equivariant graph convolutional layer. For the practical case where the perfectly regular graph does not exist, we design two quantitative measures to evaluate the degree of irregularity for a spherical graph. The Geodesic ICOsahedral Pixelation (GICOPix) is adopted to construct spherical graphs with the minimum degree of irregularity compared to the current popular pixelation schemes. In addition, we design a hierarchical pooling layer to keep the rotation-equivariance, followed by a transition layer to enforce the invariance to the rotations for spherical image classification. We evaluate the proposed graph convolutional layers with different pixelations schemes in terms of equivariance errors. We also assess the effectiveness of the proposed SGCN in fulfilling rotation-invariance by the invariance error of the transition layers and recognizing the spherical images and 3D objects.
****
Composing Good Shots by Exploiting Mutual Relations
Debang Li, Junge Zhang, Kaiqi Huang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4213-4222
Finding views with a good composition from an input image is a common but challenging problem. There are usually at least dozens of candidates (regions) in an image, and how to evaluate these candidates is subjective. Most existing methods only use the feature corresponding to each candidate to evaluate the quality. However, the mutual relations between the candidates from an image play an essenti

al role in composing a good shot due to the comparative nature of this problem. Motivated by this, we propose a graph-based module with a gated feature update to model the relations between different candidates. The candidate region features are propagated on a graph that models mutual relations between different regions for mining the useful information such that the relation features and region features are adaptively fused. We design a multi-task loss to train the model, especially, a regularization term is adopted to incorporate the prior knowledge about the relations into the graph. A data augmentation method is also developed by mixing nodes from different graphs to improve the model generalization ability. Experimental results show that the proposed model performs favorably against state-of-the-art methods, and comprehensive ablation studies demonstrate the contribution of each module and graph-based inference of the proposed method.
****

DIST: Rendering Deep Implicit Signed Distance Function With Differentiable Sphere Tracing

Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, Zhaopeng Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2019-2028

We propose a differentiable sphere tracing algorithm to bridge the gap between inverse graphics methods and the recently proposed deep learning based implicit signed distance function. Due to the nature of the implicit function, the rendering process requires tremendous function queries, which is particularly problematic when the function is represented as a neural network. We optimize both the forward and backward pass of our rendering layer to make it run efficiently with affordable memory consumption on a commodity graphics card. Our rendering method is fully differentiable such that losses can be directly computed on the rendered 2D observations, and the gradients can be propagated backward to optimize the 3D geometry. We show that our rendering method can effectively reconstruct accurate 3D shapes from various inputs, such as sparse depth and multi-view images, through inverse optimization. With the geometry based reasoning, our 3D shape prediction methods show excellent generalization capability and robustness against various noises.
****

Progressive Relation Learning for Group Activity Recognition

Guyue Hu, Bo Cui, Yuan He, Shan Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 980-989

Group activities usually involve spatio-temporal dynamics among many interactive individuals, while only a few participants at several key frames essentially define the activity. Therefore, effectively modeling the group-relevant and suppressing the irrelevant actions (and interactions) are vital for group activity recognition. In this paper, we propose a novel method based on deep reinforcement learning to progressively refine the low-level features and high-level relations of group activities. Firstly, we construct a semantic relation graph (SRG) to explicitly model the relations among persons. Then, two agents adopting policy according to two Markov decision processes are applied to progressively refine the SRG. Specifically, one feature-distilling (FD) agent in the discrete action space refines the low-level spatio-temporal features by distilling the most informative frames. Another relation-gating (RG) agent in continuous action space adjusts the high-level semantic graph to pay more attention to group-relevant relations. The SRG, FD agent, and RG agent are optimized alternately to mutually boost the performance of each other. Extensive experiments on two widely used benchmarks demonstrate the effectiveness and superiority of the proposed approach.
****

Boundary-Aware 3D Building Reconstruction From a Single Overhead Image

Jisan Mahmud, True Price, Akash Bapat, Jan-Michael Frahm; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 441-451

We propose a boundary-aware multi-task deep-learning-based framework for fast 3D building modeling from a single overhead image. Unlike most existing techniques which rely on multiple images for 3D scene modeling, we seek to model the build

ings in the scene from a single overhead image by jointly learning a modified si gned distance function (SDF) from the building boundaries, a dense heightmap of the scene, and scene semantics. To jointly train for these tasks, we leverage pi xel-wise semantic segmentation and normalized digital surface maps (nDSM) as sup ervision, in addition to labeled building outlines. At test time, buildings in t he scene are automatically modeled in 3D using only an input overhead image. We demonstrate an increase in building modeling performance using a multi-feature n etwork architecture that improves building outline detection by considering netw ork features learned for the other jointly learned tasks. We also introduce a no vel mechanism for robustly refining instance-specific building outlines using th e learned modified SDF. We verify the effectiveness of our method on multiple la rge-scale satellite and aerial imagery datasets, where we obtain state-of-the-ar t performance in the 3D building reconstruction task.
****

Learning Formation of Physically-Based Face Attributes
Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, Hao Li; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 3410-3419
Based on a combined data set of 4000 high resolution facial scans, we introduce a non-linear morphable face model, capable of producing multifarious face geomet ry of pore-level resolution, coupled with material attributes for use in physica lly-based rendering. We aim to maximize the variety of the participant's face id entities, while increasing the robustness of correspondence between unique compo nents, including middle-frequency geometry, albedo maps, specular intensity maps and high-frequency displacement details. Our deep learning based generative mod el learns to correlate albedo and geometry, which ensures the anatomical correct ness of the generated assets. We demonstrate potential use of our generative mod el for novel identity generation, model fitting, interpolation, animation, high fidelity data visualization, and low-to-high resolution data domain transferring . We hope the release of this generative model will encourage further cooperatio n between all graphics, vision, and data focused professionals, while demonstrat ing the cumulative value of every individual's complete biometric profile.
****

Deep Metric Learning via Adaptive Learnable Assessment
Wenzhao Zheng, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2020, pp. 2960-2969
In this paper, we propose a deep metric learning via adaptive learnable assessme nt (DML-ALA) method for image retrieval and clustering, which aims to learn a sa mple assessment strategy to maximize the generalization of the trained metric. U nlike existing deep metric learning methods that usually utilize a fixed samplin g strategy like hard negative mining, we propose a sequence-aware learnable asse ssor which re-weights each training example to train the metric towards good gen eralization. We formulate the learning of this assessor as a meta-learning probl em, where we employ an episode-based training scheme and update the assessor at each iteration to adapt to the current model status. We construct each episode b y sampling two subsets of disjoint labels to simulate the procedure of training and testing and use the performance of one-gradient-updated metric on the valida tion subset as the meta-objective of the assessor. Experimental results on the w idely used CUB-200-2011, Cars196, and Stanford Online Products datasets demonstr ate the effectiveness of the proposed approach.
****

Rethinking Computer-Aided Tuberculosis Diagnosis
Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 020, pp. 2646-2655
As a serious infectious disease, tuberculosis (TB) is one of the major threats t o human health worldwide, leading to millions of death every year. Although earl y diagnosis and treatment can greatly improve the chances of survival, it remain s a major challenge, especially in developing countries. Computer-aided tubercul

osis diagnosis (CTD) is a promising choice for TB diagnosis due to the great suc cesses of deep learning. However, when it comes to TB diagnosis, the lack of tra ining data has hampered the progress of CTD. To solve this problem, we establish a large-scale TB dataset, namely Tuberculosis X-ray (TBX11K) dataset. This data set contains 11200 X-ray images with corresponding bounding box annotations for TB areas, while the existing largest public TB dataset only has 662 X-ray images with corresponding image-level annotations. The proposed dataset enables the tr aining of sophisticated detectors for high-quality CTD. We reform the existing o bject detectors to adapt them to simultaneous image classification and TB area d etection. These reformed detectors are trained and evaluated on the proposed TBX 11K dataset and served as the baselines for future research.
****

Creating Something From Nothing: Unsupervised Knowledge Distillation for Cross-M odal Hashing

Hengtong Hu, Lingxi Xie, Richang Hong, Qi Tian; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3123-3132

In recent years, cross-modal hashing (CMH) has attracted increasing attentions, mainly because its potential ability of mapping contents from different modaliti es, especially in vision and language, into the same space, so that it becomes e fficient in cross-modal data retrieval. There are two main frameworks for CMH, d iffering from each other in whether semantic supervision is required. Compared t o the unsupervised methods, the supervised methods often enjoy more accurate res ults, but require much heavier labors in data annotation. In this paper, we prop ose a novel approach that enables guiding a supervised method using outputs prod uced by an unsupervised method. Specifically, we make use of teacher-student opt imization for propagating knowledge. Experiments are performed on two popular CM H benchmarks, i.e., the MIRFlickr and NUS-WIDE datasets. Our approach outperform s all existing unsupervised methods by a large margin.
****

Adversarial Examples Improve Image Recognition

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2020, pp. 819-828

Adversarial examples are commonly viewed as a threat to ConvNets. Here we presen t an opposite perspective: adversarial examples can be used to improve image rec ognition models if harnessed in the right manner. We propose AdvProp, an enhance d adversarial training scheme which treats adversarial examples as additional ex amples, to prevent overfitting. Key to our method is the usage of a separate aux iliary batch norm for adversarial examples, as they have different underlying di stributions to normal examples. We show that AdvProp improves a wide range of mo dels on various image recognition tasks and performs better when the models are bigger. For instance, by applying AdvProp to the latest EfficientNet-B7 [28] on ImageNet, we achieve significant improvements on ImageNet (+0.7%), ImageNet-C (+ 6.5%), ImageNet-A (+7.0%), Stylized-ImageNet (+4.8%). With an enhanced Efficient Net-B8, our method achieves the state-of-the-art 85.5% ImageNet top-1 accuracy w ithout extra data. This result even surpasses the best model in [20] which is tr ained with 3.5B Instagram images (3000X more than ImageNet) and 9.4X more para meters. Code and models will be made publicly available.
****

TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution

Yapeng Tian, Yulun Zhang, Yun Fu, Chenliang Xu; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3360-3369

Video super-resolution (VSR) aims to restore a photo-realistic high-resolution (HR) video frame from both its corresponding low-resolution (LR) frame (reference frame) and multiple neighboring frames (supporting frames). Due to varying moti on of cameras or objects, the reference frame and each support frame are not ali gned. Therefore, temporal alignment is a challenging yet important problem for V SR. Previous VSR methods usually utilize optical flow between the reference fram e and each supporting frame to warp the supporting frame for temporal alignment. However, both inaccurate flow and the image-level warping strategy will lead to

artifacts in the warped supporting frames. To overcome the limitation, we propose a temporally-deformable alignment network (TDAN) to adaptively align the reference frame and each supporting frame at the feature level without computing optical flow. The TDAN uses features from both the reference frame and each supporting frame to dynamically predict offsets of sampling convolution kernels. By using the corresponding kernels, TDAN transforms supporting frames to align with the reference frame. To predict the HR video frame, a reconstruction network taking aligned frames and the reference frame is utilized. Experimental results demonstrate that the TDAN is capable of alleviating occlusions and artifacts for temporal alignment and the TDAN-based VSR model outperforms several recent state-of-the-art VSR networks with a comparable or even much smaller model size. The source code and pre-trained models are released in https://github.com/YapengTian/TDAN-VSR.

****

Retina-Like Visual Image Reconstruction via Spiking Neural Model

Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1438-1446

The high-sensitivity vision of primates, including humans, is mediated by a small retinal region called the fovea. As a novel bio-inspired vision sensor, spike camera mimics the fovea to record the nature scenes by continuous-time spikes instead of frame-based manner. However, reconstructing visual images from the spikes remains to be a challenge. In this paper, we design a retina-like visual image reconstruction framework, which is flexible in reconstructing full texture of natural scenes from the totally new spike data. Specifically, the proposed architecture consists of motion local excitation layer, spike refining layer and visual reconstruction layer motivated by bio-realistic leaky integrate and fire (LIF) neurons and synapse connection with spike-timing-dependent plasticity (STDP) rules. This approach may represent a major shift from conventional frame-based vision to the continuous-time retina-like vision, owning to the advantages of high temporal resolution and low power consumption. To test the performance, a spike dataset is constructed which is recorded by the spike camera. The experimental results show that the proposed approach is extremely effective in reconstructing the visual image in both normal and high speed scenes, while achieving high dynamic range and high image quality.

****

Resolution Adaptive Networks for Efficient Inference

Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2369-2378

Adaptive inference is an effective mechanism to achieve a dynamic tradeoff between accuracy and computational cost in deep networks. Existing works mainly exploit architecture redundancy in network depth or width. In this paper, we focus on spatial redundancy of input samples and propose a novel Resolution Adaptive Network (RANet), which is inspired by the intuition that low-resolution representations are sufficient for classifying "easy" inputs containing large objects with prototypical features, while only some "hard" samples need spatially detailed information. In RANet, the input images are first routed to a lightweight sub-network that efficiently extracts low-resolution representations, and those samples with high prediction confidence will exit early from the network without being further processed. Meanwhile, high-resolution paths in the network maintain the capability to recognize the "hard" samples. Therefore, RANet can effectively reduce the spatial redundancy involved in inferring high-resolution inputs. Empirically, we demonstrate the effectiveness of the proposed RANet on the CIFAR-10, CIFAR-100 and ImageNet datasets in both the anytime prediction setting and the budgeted batch classification setting.

****

Normal Assisted Stereo Depth Estimation

Uday Kusupati, Shuo Cheng, Rui Chen, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2189-2199

Accurate stereo depth estimation plays a critical role in various 3D tasks in both indoor and outdoor environments. Recently, learning-based multi-view stereo methods have demonstrated competitive performance with limited number of views. However, in challenging scenarios, especially when building cross-view correspondences is hard, these methods still cannot produce satisfying results. In this paper, we study how to enforce the consistency between surface normal and depth at training time to improve the performance. We couple the learning of a multi-view normal estimation module and a multi-view depth estimation module. In addition, we propose a novel consistency loss to train an independent consistency module that refines the depths from depth/normal pairs. We find that the joint learning can improve both the prediction of normal and depth, and the accuracy and smoothness can be further improved by enforcing the consistency. Experiments on MVS, SUN3D, RGBD and Scenes11 demonstrate the effectiveness of our method and state-of-the-art performance.

****

Inverse Rendering for Complex Indoor Scenes: Shape, Spatially-Varying Lighting and SVBRDF From a Single Image

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2475-2484

We propose a deep inverse rendering framework for indoor scenes. From a single RGB image of an arbitrary indoor scene, we obtain a complete scene reconstruction, estimating shape, spatially-varying lighting, and spatially-varying, non-Lambertian surface reflectance. Our novel inverse rendering network incorporates physical insights -- including a spatially-varying spherical Gaussian lighting representation, a differentiable rendering layer to model scene appearance, a cascade structure to iteratively refine the predictions and a bilateral solver for refinement -- allowing us to jointly reason about shape, lighting, and reflectance. Since no existing dataset provides ground truth high quality spatially-varying material and spatially-varying lighting, we propose novel methods to map complex materials to existing indoor scene datasets and a new physically-based GPU renderer to create a large-scale, photorealistic indoor dataset. Experiments show that our framework outperforms previous methods and enables various novel applications like photorealistic object insertion and material editing.

****

Explorable Super Resolution

Yuval Bahat, Tomer Michaeli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2716-2725

Single image super resolution (SR) has seen major performance leaps in recent years. However, existing methods do not allow exploring the infinitely many plausible reconstructions that might have given rise to the observed low-resolution (LR) image. These different explanations to the LR image may dramatically vary in their textures and fine details, and may often encode completely different semantic information. In this paper, we introduce the task of explorable super resolution. We propose a framework comprising a graphical user interface with a neural network backend, allowing editing the SR output so as to explore the abundance of plausible HR explanations to the LR input. At the heart of our method is a novel module that can wrap any existing SR network, analytically guaranteeing that its SR outputs would precisely match the LR input, when down- sampled. Besides its importance in our setting, this module is guaranteed to decrease the reconstruction error of any SR network it wraps, and can be used to cope with blur kernels that are different from the one the network was trained for. We illustrate our approach in a variety of use cases, ranging from medical imaging and forensics, to graphics.

****

DuDoRNet: Learning a Dual-Domain Recurrent Network for Fast MRI Reconstruction With Deep T1 Prior

Bo Zhou, S. Kevin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4273-4282

MRI with multiple protocols is commonly used for diagnosis, but it suffers from

a long acquisition time, which yields the image quality vulnerable to say motion artifacts. To accelerate, various methods have been proposed to reconstruct full images from under-sampled k-space data. However, these algorithms are inadequate for two main reasons. Firstly, aliasing artifacts generated in the image domain are structural and non-local, so that sole image domain restoration is insufficient. Secondly, though MRI comprises multiple protocols during one exam, almost all previous studies only employ the reconstruction of an individual protocol using a highly distorted undersampled image as input, leaving the use of fully-sampled short protocol (say T1) as complementary information highly underexplored. In this work, we address the above two limitations by proposing a Dual Domain Recurrent Network (DuDoRNet) with deep T1 prior embedded to simultaneously recover k-space and images for accelerating the acquisition of MRI with a long imaging protocol. Specifically, a Dilated Residual Dense Network (DRDNet) is customized for dual domain restorations from undersampled MRI data. Extensive experiments on different sampling patterns and acceleration rates demonstrate that our method consistently outperforms state-of-the-art methods, and can reconstruct high quality MRI.

****

Unsupervised Instance Segmentation in Microscopy Images via Panoptic Domain Adaptation and Task Re-Weighting

Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, Weidong Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4243-4252

Unsupervised domain adaptation (UDA) for nuclei instance segmentation is important for digital pathology, as it alleviates the burden of labor-intensive annotation and domain shift across datasets. In this work, we propose a Cycle Consistency Panoptic Domain Adaptive Mask R-CNN (CyC-PDAM) architecture for unsupervised nuclei segmentation in histopathology images, by learning from fluorescence microscopy images. More specifically, we first propose a nuclei inpainting mechanism to remove the auxiliary generated objects in the synthesized images. Secondly, a semantic branch with a domain discriminator is designed to achieve panoptic-level domain adaptation. Thirdly, in order to avoid the influence of the source-biased features, we propose a task re-weighting mechanism to dynamically add trade-off weights for the task-specific loss functions. Experimental results on three datasets indicate that our proposed method outperforms state-of-the-art UDA methods significantly, and demonstrates a similar performance as fully supervised methods.

****

Heterogeneous Knowledge Distillation Using Information Flow Modeling

Nikolaos Passalis, Maria Tzelepi, Anastasios Tefas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2339-2348

Knowledge Distillation (KD) methods are capable of transferring the knowledge encoded in a large and complex teacher into a smaller and faster student. Early methods were usually limited to transferring the knowledge only between the last layers of the networks, while latter approaches were capable of performing multi-layer KD, further increasing the accuracy of the student. However, despite their improved performance, these methods still suffer from several limitations that restrict both their efficiency and flexibility. First, existing KD methods typically ignore that neural networks undergo through different learning phases during the training process, which often requires different types of supervision for each one. Furthermore, existing multi-layer KD methods are usually unable to effectively handle networks with significantly different architectures (heterogeneous KD). In this paper we propose a novel KD method that works by modeling the information flow through the various layers of the teacher model and then training a student model to mimic this information flow. The proposed method is capable of overcoming the aforementioned limitations by using an appropriate supervision scheme during the different phases of the training process, as well as by designing and training an appropriate auxiliary teacher model that acts as a proxy model capable of "explaining" the way the teacher works to the student. The effect

iveness of the proposed method is demonstrated using four image datasets and several different evaluation setups.

****

CARS: Continuous Evolution for Efficient Neural Architecture Search

Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1829-1838

Searching techniques in most of existing neural architecture search (NAS) algorithms are mainly dominated by differentiable methods for the efficiency reason. In contrast, we develop an efficient continuous evolutionary approach for searching neural networks. Architectures in the population that share parameters within one SuperNet in the latest generation will be tuned over the training dataset with a few epochs. The searching in the next evolution generation will directly inherit both the SuperNet and the population, which accelerates the optimal network generation. The non-dominated sorting strategy is further applied to preserve only results on the Pareto front for accurately updating the SuperNet. Several neural networks with different model sizes and performances will be produced after the continuous search with only 0.4 GPU days. As a result, our framework provides a series of networks with the number of parameters ranging from 3.7M to 5.1M under mobile settings. These networks surpass those produced by the state-of-the-art methods on the benchmark ImageNet dataset.

****

Bi-Directional Interaction Network for Person Search

Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2839-2848

Existing works have designed end-to-end frameworks based on Faster-RCNN for person search. Due to the large receptive fields in deep networks, the feature maps of each proposal, cropped from the stem feature maps, involve redundant context information outside the bounding boxes. However, person search is a fine-grained task which needs accurate appearance information. Such context information can make the model fail to focus on persons, so the learned representations lack the capacity to discriminate various identities. To address this issue, we propose a Siamese network which owns an additional instance-aware branch, named Bi-directional Interaction Network (BINet). During the training phase, in addition to scene images, BINet also takes as inputs person patches which help the model discriminate identities based on human appearance. Moreover, two interaction losses are designed to achieve bi-directional interaction between branches at two levels. The interaction can help the model learn more discriminative features for persons in the scene. At the inference stage, only the major branch is applied, so BINet introduces no additional computation. Extensive experiments on two widely used person search benchmarks, CUHK-SYSU and PRW, have shown that our BINet achieves state-of-the-art results among end-to-end methods without loss of efficiency.

****

ZSTAD: Zero-Shot Temporal Activity Detection

Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, Alexander Hauptmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 879-888

An integral part of video analysis and surveillance is temporal activity detection, which means to simultaneously recognize and localize activities in long untrimmed videos. Currently, the most effective methods of temporal activity detection are based on deep learning, and they typically perform very well with large scale annotated videos for training. However, these methods are limited in real applications due to the unavailable videos about certain activity classes and the time-consuming data annotation. To solve this challenging problem, we propose a novel task setting called zero-shot temporal activity detection (ZSTAD), where activities that have never been seen in training can still be detected. We design an end-to-end deep network based on R-C3D as the architecture for this solution. The proposed network is optimized with an innovative loss function that consi

ders the embeddings of activity labels and their super-classes while learning th e common semantics of seen and unseen activities. Experiments on both the THUMOS '14 and the Charades datasets show promising performance in terms of detecting u nseen activities.

****

Unpaired Image Super-Resolution Using Pseudo-Supervision

Shunta Maeda; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2020, pp. 291-300

In most studies on learning-based image super-resolution (SR), the paired traini ng dataset is created by downscaling high-resolution (HR) images with a predeter mined operation (e.g., bicubic). However, these methods fail to super-resolve re al-world low-resolution (LR) images, for which the degradation process is much m ore complicated and unknown. In this paper, we propose an unpaired SR method usi ng a generative adversarial network that does not require a paired/aligned train ing dataset. Our network consists of an unpaired kernel/noise correction network and a pseudo-paired SR network. The correction network removes noise and adjust s the kernel of the inputted LR image; then, the corrected clean LR image is ups caled by the SR network. In the training phase, the correction network also prod uces a pseudo-clean LR image from the inputted HR image, and then a mapping from the pseudo-clean LR image to the inputted HR image is learned by the SR network in a paired manner. Because our SR network is independent of the correction net work, well-studied existing network architectures and pixel-wise loss functions can be integrated with the proposed framework. Experiments on diverse datasets s how that the proposed method is superior to existing solutions to the unpaired S R problem.

****

Training Quantized Neural Networks With a Full-Precision Auxiliary Module

Bohan Zhuang, Lingqiao Liu, Mingkui Tan, Chunhua Shen, Ian Reid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 020, pp. 1488-1497

In this paper, we seek to tackle a challenge in training low-precision networks: the notorious difficulty in propagating gradient through a low-precision networ k due to the non-differentiable quantization function. We propose a solution by training the low-precision network with a full-precision auxiliary module. Speci fically, during training, we construct a mix-precision network by augmenting the original low-precision network with the full precision auxiliary module. Then t he augmented mix-precision network and the low-precision network are jointly opt imized. This strategy creates additional full-precision routes to update the par ameters of the low-precision model, thus making the gradient back-propagates mor e easily. At the inference time, we discard the auxiliary module without introdu cing any computational complexity to the low-precision network. We evaluate the proposed method on image classification and object detection over various quanti zation approaches and show consistent performance increase. In particular, we ac hieve near lossless performance to the full-precision model by using a 4-bit det ector, which is of great practical value.

****

ReSprop: Reuse Sparsified Backpropagation

Negar Goli, Tor M. Aamodt; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2020, pp. 1548-1558

The success of Convolutional Neural Networks (CNNs) in various applications is a ccompanied by a significant increase in computation and training time. In this w ork, we focus on accelerating training by observing that about 90% of gradients are reusable during training. Leveraging this observation, we propose a new algo rithm, Reuse-Sparse-Backprop (ReSprop), as a method to sparsify gradient vectors during CNN training. ReSprop maintains state-of-the-art accuracy on CIFAR-10, C IFAR-100, and ImageNet datasets with less than 1.1% accuracy loss while enabling a reduction in back-propagation computations by a factor of 10x resulting in a 2.7x overall speedup in training. As the computation reduction introduced by Re-Sprop is accomplished by introducing fine-grained sparsity that reduces computat ion efficiency on GPUs, we introduce a generic sparse convolution neural network

accelerator (GSCN), which is designed to accelerate sparse back-propagation convolutions. When combined with ReSprop, GSCN achieves 8.0x and 7.2x speedup in the backward pass on ResNet34 and VGG16 versus a GTX 1080 Ti GPU.
****

Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network
Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3667-3676
Blind image quality assessment (BIQA) for authentically distorted images has always been a challenging problem, since images captured in the wild include varies contents and diverse types of distortions. The vast majority of prior BIQA methods focus on how to predict synthetic image quality, but fail when applied to real-world distorted images. To deal with the challenge, we propose a self-adaptive hyper network architecture to blind assess image quality in the wild. We separate the IQA procedure into three stages including content understanding, perception rule learning and quality predicting. After extracting image semantics, perception rule is established adaptively by a hyper network, and then adopted by a quality prediction network. In our model, image quality can be estimated in a self-adaptive manner, thus generalizes well on diverse images captured in the wild. Experimental results verify that our approach not only outperforms the state-of-the-art methods on challenging authentic image databases but also achieves competing performances on synthetic image databases, though it is not explicitly designed for the synthetic task.
****

Bundle Adjustment on a Graph Processor
Joseph Ortiz, Mark Pupilli, Stefan Leutenegger, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2416-2425
Graph processors such as Graphcore's Intelligence Processing Unit (IPU) are part of the major new wave of novel computer architecture for AI, and have a general design with massively parallel computation, distributed on-chip memory and very high inter-core communication bandwidth which allows breakthrough performance for message passing algorithms on arbitrary graphs. We show for the first time that the classical computer vision problem of bundle adjustment (BA) can be solved extremely fast on a graph processor using Gaussian Belief Propagation. Our simple but fully parallel implementation uses the 1216 cores on a single IPU chip to, for instance, solve a real BA problem with 125 keyframes and 1919 points in under 40ms, compared to 1450ms for the Ceres CPU library. Further code optimisation will surely increase this difference on static problems, but we argue that the real promise of graph processing is for flexible in-place optimisation of general, dynamically changing factor graphs representing Spatial AI problems. We give indications of this with experiments showing the ability of GBP to efficiently solve incremental SLAM problems, and deal with robust cost functions and different types of factors.
****

Multi-View Neural Human Rendering
Minye Wu, Yuehao Wang, Qiang Hu, Jingyi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1682-1691
We present an end-to-end Neural Human Renderer (NHR) for dynamic human captures under the multi-view setting. NHR adopts PointNet++ for feature extraction (FE) to enable robust 3D correspondence matching on low quality, dynamic 3D reconstructions. To render new views, we map 3D features onto the target camera as a 2D feature map and employ an anti-aliased CNN to handle holes and noises. Newly synthesized views from NHR can be further used to construct visual hulls to handle textureless and/or dark regions such as black clothing. Comprehensive experiments show NHR significantly outperforms the state-of-the-art neural and image-based rendering techniques, especially on hands, hair, nose, foot, etc.
****

Self-Supervised Monocular Trained Depth Estimation Using Self-Attention and Discrete Disparity Volume

Adrian Johnston, Gustavo Carneiro; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2020, pp. 4756-4765
Monocular depth estimation has become one of the most studied applications in co
mputer vision, where the most accurate approaches are based on fully supervised
 learning models. However, the acquisition of accurate and large ground truth dat
a sets to model these fully supervised methods is a major challenge for the furt
her development of the area. Self-supervised methods trained with monocular vide
os constitute one the most promising approaches to mitigate the challenge mentio
ned above due to the wide-spread availability of training data. Consequently, th
ey have been intensively studied, where the main ideas explored consist of diffe
rent types of model architectures, loss functions, and occlusion masks to addres
s non-rigid motion. In this paper, we propose two new ideas to improve self-supe
rvised monocular trained depth estimation: 1) self-attention, and 2) discrete di
sparity prediction. Compared with the usual localised convolution operation, sel
f-attention can explore a more general contextual information that allows the in
ference of similar disparity values at non-contiguous regions of the image. Disc
rete disparity prediction has been shown by fully supervised methods to provide
 a more robust and sharper depth estimation than the more common continuous dispa
rity prediction, besides enabling the estimation of depth uncertainty. We show t
hat the extension of the state-of-the-art self-supervised monocular trained dept
h estimator Monodepth2 with these two ideas allows us to design a model that pro
duces the best results in the field in KITTI 2015 and Make3D, closing the gap wi
th respect self-supervised stereo training and fully supervised approaches.
****

Joint Filtering of Intensity Images and Neuromorphic Events for High-Resolution
Noise-Robust Imaging
Zihao W. Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huan
g, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2020, pp. 1609-1619
We present a novel computational imaging system with high resolution and low noi
se. Our system consists of a traditional video camera which captures high-resolu
tion intensity images, and an event camera which encodes high-speed motion as a
 stream of asynchronous binary events. To process the hybrid input, we propose a
 unifying framework that first bridges the two sensing modalities via a noise-rob
ust motion compensation model, and then performs joint image filtering. The filt
ered output represents the temporal gradient of the captured space-time volume,
which can be viewed as motion-compensated event frames with high resolution and
low noise. Therefore, the output can be widely applied to many existing event-ba
sed algorithms that are highly dependent on spatial resolution and noise robustn
ess. In experimental results performed on both publicly available datasets as we
ll as our contributing RGB-DAVIS dataset, we show systematic performance improve
ment in applications such as high frame-rate video synthesis, feature/corner det
ection and tracking, as well as high dynamic range image reconstruction.
****

Automatic Neural Network Compression by Sparsity-Quantization Joint Learning: A
Constrained Optimization-Based Approach
Haichuan Yang, Shupeng Gui, Yuhao Zhu, Ji Liu; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2178-2188
Deep Neural Networks (DNNs) are applied in a wide range of usecases. There is an
 increased demand for deploying DNNs on devices that do not have abundant resour
ces such as memory and computation units. Recently, network compression through
a variety of techniques such as pruning and quantization have been proposed to r
educe the resource requirement. A key parameter that all existing compression te
chniques are sensitive to is the compression ratio (e.g., pruning sparsity, quan
tization bitwidth) of each layer. Traditional solutions treat the compression ra
tios of each layer as hyper-parameters, and tune them using human heuristic. Rec
ent researchers start using black-box hyper-parameter optimizations, but they wi
ll introduce new hyper-parameters and have efficiency issue. In this paper, we p
ropose a framework to jointly prune and quantize the DNNs automatically accordin
g to a target model size without using any hyper-parameters to manually set the

compression ratio for each layer. In the experiments, we show that our framework can compress the weights data of ResNet-50 to be 836x smaller without accuracy loss on CIFAR-10, and compress AlexNet to be 205x smaller without accuracy loss on ImageNet classification.
****

When2com: Multi-Agent Perception via Communication Graph Grouping
Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4106-4115
While significant advances have been made for single-agent perception, many applications require multiple sensing agents and cross-agent communication due to benefits such as coverage and robustness. It is therefore critical to develop frameworks which support multi-agent collaborative perception in a distributed and bandwidth-efficient manner. In this paper, we address the collaborative perception problem, where one agent is required to perform a perception task and can communicate and share information with other agents on the same task. Specifically, we propose a communication framework by learning both to construct communication groups and decide when to communicate. We demonstrate the generalizability of our framework on two different perception tasks and show that it significantly reduces communication bandwidth while maintaining superior performance.
****

MAGSAC++, a Fast, Reliable and Accurate Robust Estimator
Daniel Barath, Jana Noskova, Maksym Ivashechkin, Jiri Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1304-1312
We propose MAGSAC++ and Progressive NAPSAC sampler, P-NAPSAC in short. In MAGSAC++, we replace the model quality and polishing functions of the original method by an iteratively re-weighted least-squares fitting with weights determined via marginalizing over the noise scale. MAGSAC++ is fast -- often an order of magnitude faster -- and more geometrically accurate than MAGSAC. P-NAPSAC merges the advantages of local and global sampling by drawing samples from gradually growing neighborhoods. Exploiting that nearby points are more likely to originate from the same geometric model, P-NAPSAC finds local structures earlier than global samplers. We show that the progressive spatial sampling in P-NAPSAC can be integrated with PROSAC sampling, which is applied to the first, location-defining, point. The methods are tested on homography and fundamental matrix fitting on six publicly available datasets. MAGSAC combined with P-NAPSAC sampler is superior to state-of-the-art robust estimators in terms of speed, accuracy and failure rate.
****

Organ at Risk Segmentation for Head and Neck Cancer Using Stratified Learning and Neural Architecture Search
Dazhou Guo, Dakai Jin, Zhuotun Zhu, Tsung-Ying Ho, Adam P. Harrison, Chun-Hung Chao, Jing Xiao, Le Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4223-4232
OAR segmentation is a critical step in radiotherapy of head and neck (H&N) cancer, where inconsistencies across radiation oncologists and prohibitive labor costs motivate automated approaches. However, leading methods using standard fully convolutional network workflows that are challenged when the number of OARs becomes large, e.g. > 40. For such scenarios, insights can be gained from the stratification approaches seen in manual clinical OAR delineation. This is the goal of our work, where we introduce stratified organ at risk segmentation (SOARS), an approach that stratifies OARs into anchor, mid-level, and small & hard (S&H) categories. SOARS stratifies across two dimensions. The first dimension is that distinct processing pipelines are used for each OAR category. In particular, inspired by clinical practices, anchor OARs are used to guide the mid-level and S&H categories. The second dimension is that distinct network architectures are used to manage the significant contrast, size, and anatomy variations between different OARs. We use differentiable neural architecture search (NAS), allowing the network to choose among 2D, 3D or Pseudo-3D convolutions. Extensive 4-fold cross-validation on 142 H&N cancer patients with 42 manually labeled OARs, the most compr

ehensive OAR dataset to date, demonstrates that both pipeline- and NAS-stratification significantly improves quantitative performance over the state-of-the-art (from 69.52% to 73.68% in absolute Dice scores). Thus, SOARS provides a powerful and principled means to manage the highly complex segmentation space of OARs.
****

Learning Human-Object Interaction Detection Using Interaction Points
Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4116-4125
Understanding interactions between humans and objects is one of the fundamental problems in visual classification and an essential step towards detailed scene understanding. Human-object interaction (HOI) detection strives to localize both the human and an object as well as the identification of complex interactions between them. Most existing HOI detection approaches are instance-centric where interactions between all possible human-object pairs are predicted based on appearance features and coarse spatial information. We argue that appearance features alone are insufficient to capture complex human-object interactions. In this paper, we therefore propose a novel fully-convolutional approach that directly detects the interactions between human-object pairs. Our network predicts interaction points, which directly localize and classify the inter-action. Paired with the densely predicted interaction vectors, the interactions are associated with human and object detections to obtain final predictions. To the best of our knowledge, we are the first to propose an approach where HOI detection is posed as a keypoint detection and grouping problem. Experiments are performed on two popular benchmarks: V-COCO and HICO-DET. Our approach sets a new state-of-the-art on both datasets. Code is available at https://github.com/vaesl/IP-Net.
****

Deep Kinematics Analysis for Monocular 3D Human Pose Estimation
Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 899-908
For monocular 3D pose estimation conditioned on 2D detection, noisy/unreliable input is a key obstacle in this task. Simple structure constraints attempting to tackle this problem, e.g., symmetry loss and joint angle limit, could only provide marginal improvements and are commonly treated as auxiliary losses in previous researches. Thus it still remains challenging about how to effectively utilize the power of human prior knowledge for this task. In this paper, we propose to address above issue in a systematic view. Firstly, we show that optimizing the kinematics structure of noisy 2D inputs is critical to obtain accurate 3D estimations. Secondly, based on corrected 2D joints, we further explicitly decompose articulated motion with human topology, which leads to more compact 3D static structure easier for estimation. Finally, temporal refinement emphasizing the validity of 3D dynamic structure is naturally developed to pursue more accurate result. Above three steps are seamlessly integrated into deep neural models, which form a deep kinematics analysis pipeline concurrently considering the static/dynamic structure of 2D inputs and 3D outputs. Extensive experiments show that proposed framework achieves state-of-the-art performance on two widely used 3D human action datasets. Meanwhile, targeted ablation study shows that each former step is critical for the latter one to obtain promising results.
****

Domain Decluttering: Simplifying Images to Mitigate Synthetic-Real Domain Shift and Improve Depth Estimation
Yunhan Zhao, Shu Kong, Daeyun Shin, Charless Fowlkes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3330-3340
Leveraging synthetically rendered data offers great potential to improve monocular depth estimation and other geometric estimation tasks, but closing the synthetic-real domain gap is a non-trivial and important task. While much recent work has focused on unsupervised domain adaptation, we consider a more realistic scenario where a large amount of synthetic training data is supplemented by a small

set of real images with ground-truth. In this setting, we find that existing dom
ain translation approaches are difficult to train and offer little advantage ove
r simple baselines that use a mix of real and synthetic data. A key failure mode
 is that real-world images contain novel objects and clutter not present in synt
hetic training. This high-level domain shift isn't handled by existing image tra
nslation models. Based on these observations, we develop an attention module tha
t learns to identify and remove difficult out-of-domain regions in real images i
n order to improve depth prediction for a model trained primarily on synthetic d
ata. We carry out extensive experiments to validate our attend-remove-complete a
pproach (ARC) and find that it significantly outperforms state-of-the-art domain
 adaptation methods for depth prediction. Visualizing the removed regions provid
es interpretable insights into the synthetic-real domain gap.
****

End-to-End Illuminant Estimation Based on Deep Metric Learning
Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, Guoping Qiu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p
p. 3616-3625
Previous deep learning approaches to color constancy usually directly estimate i
lluminant value from input image. Such approaches might suffer heavily from bein
g sensitive to the variation of image content. To overcome this problem, we intr
oduce a deep metric learning approach named Illuminant-Guided Triplet Network (I
GTN) to color constancy. IGTN generates an Illuminant Consistent and Discriminat
ive Feature (ICDF) for achieving robust and accurate illuminant color estimation
. ICDF is composed of semantic and color features based on a learnable color his
togram scheme. In the ICDF space, regardless of the similarities of their conten
ts, images taken under the same or similar illuminants are placed close to each
other and at the same time images taken under different illuminants are placed f
ar apart. We also adopt an end-to-end training strategy to simultaneously group
image features and estimate illuminant value, and thus our approach does not hav
e to classify illuminant in a separate module. We evaluate our method on two pub
lic datasets and demonstrate our method outperforms state-of-the-art approaches.
 Furthermore, we demonstrate that our method is less sensitive to image appearan
ces, and can achieve more robust and consistent results than other methods on a
High Dynamic Range dataset.
****

PatchVAE: Learning Local Latent Codes for Recognition
Kamal Gupta, Saurabh Singh, Abhinav Shrivastava; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4746-4755
Unsupervised representation learning holds the promise of exploiting large amoun
ts of unlabeled data to learn general representations. A promising technique for
 unsupervised learning is the framework of Variational Auto-encoders (VAEs). How
ever, unsupervised representations learned by VAEs are significantly outperforme
d by those learned by supervised learning for recognition. Our hypothesis is tha
t to learn useful representations for recognition the model needs to be encourag
ed to learn about repeating and consistent patterns in data. Drawing inspiration
 from the mid-level representation discovery work, we propose PatchVAE, that rea
sons about images at patch level. Our key contribution is a bottleneck formulati
on that encourages mid-level style representations in the VAE framework. Our exp
eriments demonstrate that representations learned by our method perform much bet
ter on the recognition tasks compared to those learned by vanilla VAEs.
****

FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation
Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, Chi-Keung Tang; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2020, pp. 2869-2878
Over the past few years, we have witnessed the success of deep learning in image
 recognition thanks to the availability of large-scale human-annotated datasets
such as PASCAL VOC, ImageNet, and COCO. Although these datasets have covered a w
ide range of object categories, there are still a significant number of objects
that are not included. Can we perform the same task without a lot of human annot

ations? In this paper, we are interested in few-shot object segmentation where t
he number of annotated training examples are limited to 5 only. To evaluate and
validate the performance of our approach, we have built a few-shot segmentation
dataset, FSS-1000, which consists of 1000 object classes with pixelwise annotati
on of ground-truth segmentation. Unique in FSS-1000, our dataset contains signif
icant number of objects that have never been seen or annotated in previous datas
ets, such as tiny daily objects, merchandise, cartoon characters, logos, etc. We
 build our baseline model using standard backbone networks such as VGG-16, ResNe
t-101, and Inception. To our surprise, we found that training our model from scr
atch using FSS-1000 achieves comparable and even better results than training wi
th weights pre-trained by ImageNet which is more than 100 times larger than FSS-
1000. Both our approach and dataset are simple, effective, and easily extensible
 to learn segmentation of new object classes given very few annotated training e
xamples. Dataset is available at https://github.com/HKUSTCV/FSS-1000
****

Correction Filter for Single Image Super-Resolution: Robustifying Off-the-Shelf
Deep Super-Resolvers
Shady Abu Hussein, Tom Tirer, Raja Giryes; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1428-1437
The single image super-resolution task is one of the most examined inverse probl
ems in the past decade. In the recent years, Deep Neural Networks (DNNs) have sh
own superior performance over alternative methods when the acquisition process u
ses a fixed known downscaling kernel---typically a bicubic kernel. However, seve
ral recent works have shown that in practical scenarios, where the test data mis
match the training data (e.g. when the downscaling kernel is not the bicubic ker
nel or is not available at training), the leading DNN methods suffer from a huge
 performance drop. Inspired by the literature on generalized sampling, in this w
ork we propose a method for improving the performance of DNNs that have been tra
ined with a fixed kernel on observations acquired by other kernels. For a known
kernel, we design a closed-form correction filter that modifies the low-resoluti
on image to match one which is obtained by another kernel (e.g. bicubic), and th
us improves the results of existing pre-trained DNNs. For an unknown kernel, we
extend this idea and propose an algorithm for blind estimation of the required c
orrection filter. We show that our approach outperforms other super-resolution m
ethods, which are designed for general downscaling kernels.
****

Adversarial Robustness: From Self-Supervised Pre-Training to Fine-Tuning
Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, Zhangyang Wang
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2020, pp. 699-708
Pretrained models from self-supervision are prevalently used in fine-tuning down
stream tasks faster or for better accuracy. However, gaining robustness from pre
training is left unexplored. We introduce adversarial training into self-supervi
sion, to provide general-purpose robust pretrained models for the first time. We
 find these robust pretrained models can benefit the subsequent fine-tuning in t
wo ways: i) boosting final model robustness; ii) saving the computation cost, if
 proceeding towards adversarial fine-tuning. We conduct extensive experiments to
 demonstrate that the proposed framework achieves large performance margins (eg,
 3.83% on robust accuracy and 1.3% on standard accuracy, on the CIFAR-10 dataset
), compared with the conventional end-to-end adversarial training baseline. More
over, we find that different self-supervised pretrained models have diverse adve
rsarial vulnerability. It inspires us to ensemble several pretraining tasks, whi
ch boosts robustness more. Our ensemble strategy contributes to a further improv
ement of 3.59% on robust accuracy, while maintaining a slightly higher standard
accuracy on CIFAR-10. Our codes are available at https://github.com/TAMU-VITA/Ad
v-SS-Pretraining.
****

Efficient Adversarial Training With Transferable Adversarial Examples
Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, Atul Prakash; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

, 2020, pp. 1181-1190
Adversarial training is an effective defense method to protect classification models against adversarial attacks. However, one limitation of this approach is that it can require orders of magnitude additional training time due to high cost of generating strong adversarial examples during training. In this paper, we first show that there is high transferability between models from neighboring epochs in the same training process, i.e., adversarial examples from one epoch continue to be adversarial in subsequent epochs. Leveraging this property, we propose a novel method, Adversarial Training with Transferable Adversarial Examples (ATTA), that can enhance the robustness of trained models and greatly improve the training efficiency by accumulating adversarial perturbations through epochs. Compared to state-of-the-art adversarial training methods, ATTA enhances adversarial accuracy by up to 7.2% on CIFAR10 and requires 12 14x less training time on MNIST and CIFAR10 datasets with comparable model robustness.
****
Adversarial Texture Optimization From RGB-D Scans
Jingwei Huang, Justus Thies, Angela Dai, Abhijit Kundu, Chiyu "Max" Jiang, Leonidas J. Guibas, Matthias Niessner, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1559-1568
Realistic color texture generation is an important step in RGB-D surface reconstruction, but remains challenging in practice due to inaccuracies in reconstructed geometry, misaligned camera poses, and view-dependent imaging artifacts. In this work, we present a novel approach for color texture generation using a conditional adversarial loss obtained from weakly-supervised views. Specifically, we propose an approach to produce photorealistic textures for approximate surfaces, even from misaligned images, by learning an objective function that is robust to these errors. The key idea of our approach is to learn a patch-based conditional discriminator which guides the texture optimization to be tolerant to misalignments. Our discriminator takes a synthesized view and a real image, and evaluates whether the synthesized one is realistic, under a broadened definition of realism. We train the discriminator by providing as `real' examples pairs of input views and their misaligned versions -- so that the learned adversarial loss will tolerate errors from the scans. Experiments on synthetic and real data under quantitative or qualitative evaluation demonstrate the advantage of our approach in comparison to state of the art.
****
PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization
Shunsuke Saito, Tomas Simon, Jason Saragih, Hanbyul Joo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 84-93
Recent advances in image-based 3D human shape estimation have been driven by the significant improvement in representation power afforded by deep neural networks. Although current approaches have demonstrated the potential in real world settings, they still fail to produce reconstructions with the level of detail often present in the input images. We argue that this limitation stems primarily form two conflicting requirements; accurate predictions require large context, but precise predictions require high resolution. Due to memory limitations in current hardware, previous approaches tend to take low resolution images as input to cover large spatial context, and produce less precise (or low resolution) 3D estimates as a result. We address this limitation by formulating a multi-level architecture that is end-to-end trainable. A coarse level observes the whole image at lower resolution and focuses on holistic reasoning. This provides context to an fine level which estimates highly detailed geometry by observing higher-resolution images. We demonstrate that our approach significantly outperforms existing state-of-the-art techniques on single image human shape reconstruction by fully leveraging 1k-resolution input images.
****
TextureFusion: High-Quality Texture Acquisition for Real-Time RGB-D Scanning

Joo Ho Lee, Hyunho Ha, Yue Dong, Xin Tong, Min H. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1272-1280

Real-time RGB-D scanning technique has become widely used to progressively scan objects with a hand-held sensor. Existing online methods restore color information per voxel, and thus their quality is often limited by the tradeoff between spatial resolution and time performance. Also, such methods often suffer from blurred artifacts in the captured texture. Traditional offline texture mapping methods with non-rigid warping assume that the reconstructed geometry and all input views are obtained in advance, and the optimization takes a long time to compute mesh parameterization and warp parameters, which prevents them from being used in real-time applications. In this work, we propose a progressive texture-fusion method specially designed for real-time RGB-D scanning. To this end, we first devise a novel texture-tile voxel grid, where texture tiles are embedded in the voxel grid of the signed distance function, allowing for high-resolution texture mapping on the low-resolution geometry volume. Instead of using expensive mesh parameterization, we associate vertices of implicit geometry directly with texture coordinates. Second, we introduce real-time texture warping that applies a spatially-varying perspective mapping to input images so that texture warping efficiently mitigates the mismatch between the intermediate geometry and the current input view. It allows us to enhance the quality of texture over time while updating the geometry in real-time. The results demonstrate that the quality of our real-time texture mapping is highly competitive to that of exhaustive offline texture warping methods. Our method is also capable of being integrated into existing RGB-D scanning frameworks.

****

TomoFluid: Reconstructing Dynamic Fluid From Sparse View Videos

Guangming Zang, Ramzi Idoughi, Congli Wang, Anthony Bennett, Jianguo Du, Scott Skeen, William L. Roberts, Peter Wonka, Wolfgang Heidrich; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1870-1879

Visible light tomography is a promising and increasingly popular technique for fluid imaging. However, the use of a sparse number of viewpoints in the capturing setups makes the reconstruction of fluid flows very challenging. In this paper, we present a state-of-the-art 4D tomographic reconstruction framework that integrates several regularizers into a multi-scale matrix free optimization algorithm. In addition to existing regularizers, we propose two new regularizers for improved results: a regularizer based on view interpolation of projected images and a regularizer to encourage reprojection consistency. We demonstrate our method with extensive experiments on both simulated and real data.

****

Point Cloud Completion by Skip-Attention Network With Hierarchical Folding

Xin Wen, Tianyang Li, Zhizhong Han, Yu-Shen Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1939-1948

Point cloud completion aims to infer the complete geometries for missing regions of 3D objects from incomplete ones. Previous methods usually predict the complete point cloud based on the global shape representation extracted from the incomplete input. However, the global representation often suffers from the information loss of structure details on local regions of incomplete point cloud. To address this problem, we propose Skip-Attention Network (SA-Net) for 3D point cloud completion. Our main contributions lie in the following two-folds. First, we propose a skip-attention mechanism to effectively exploit the local structure details of incomplete point clouds during the inference of missing parts. The skip-attention mechanism selectively conveys geometric information from the local regions of incomplete point clouds for the generation of complete ones at different resolutions, where the skip-attention reveals the completion process in an interpretable way. Second, in order to fully utilize the selected geometric information encoded by skip-attention mechanism at different resolutions, we propose a novel structure-preserving decoder with hierarchical folding for complete shape gen

eration. The hierarchical folding preserves the structure of complete point clou
d generated in upper layer by progressively detailing the local regions, using t
he skip-attentioned geometry at the same resolution. We conduct comprehensive ex
periments on ShapeNet and KITTI datasets, which demonstrate that the proposed SA
-Net outperforms the state-of-the-art point cloud completion methods.
****
Revisiting Knowledge Distillation via Label Smoothing Regularization
Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, Jiashi Feng; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p
p. 3903-3911
Knowledge Distillation (KD) aims to distill the knowledge of a cumbersome teache
r model into a lightweight student model. Its success is generally attributed to
 the privileged information on similarities among categories provided by the tea
cher model, and in this sense, only strong teacher models are deployed to teach
weaker students in practice. In this work, we challenge this common belief by fo
llowing experimental observations: 1) beyond the acknowledgment that the teacher
 can improve the student, the student can also enhance the teacher significantly
 by reversing the KD procedure; 2) a poorly-trained teacher with much lower accu
racy than the student can still improve the latter significantly. To explain the
se observations, we provide a theoretical analysis of the relationships between
KD and label smoothing regularization. We prove that 1) KD is a type of learned
label smoothing regularization and 2) label smoothing regularization provides a
virtual teacher model for KD. From these results, we argue that the success of K
D is not fully due to the similarity information between categories from teacher
s, but also to the regularization of soft targets, which is equally or even more
 important. Based on these analyses, we further propose a novel Teacher-free Kno
wledge Distillation (Tf-KD) framework, where a student model learns from itself
or manually-designed regularization distribution. The Tf-KD achieves comparable
performance with normal KD from a superior teacher, which is well applied when a
 stronger teacher model is unavailable. Meanwhile, Tf-KD is generic and can be d
irectly deployed for training deep neural networks. Without any extra computatio
n cost, Tf-KD achieves up to 0.65% improvement on ImageNet over well-established
 baseline models, which is superior to label smoothing regularization.
****
Modeling Biological Immunity to Adversarial Examples
Edward Kim, Jocelyn Rego, Yijing Watkins, Garrett T. Kenyon; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020,
pp. 4666-4675
While deep learning continues to permeate through all fields of signal processin
g and machine learning, a critical exploit in these frameworks exists and remain
s unsolved. These exploits, or adversarial examples, are a type of signal attack
 that can change the output class of a classifier by perturbing the stimulus sig
nal by an imperceptible amount. The attack takes advantage of statistical irregu
larities within the training data, where the added perturbations can move the im
age across deep learning decision boundaries. What is even more alarming is the
transferability of these attacks to different deep learning models and architect
ures. This means a successful attack on one model has adversarial effects on oth
er, unrelated models. In a general sense, adversarial attack through perturbatio
ns is not a machine learning vulnerability. Human and biological vision can also
 be fooled by various methods, i.e. mixing high and low frequency images togethe
r, by altering semantically related signals, or by sufficiently distorting the i
nput signal. However, the amount and magnitude of such a distortion required to
alter biological perception is at a much larger scale. In this work, we explored
 this gap through the lens of biology and neuroscience in order to understand th
e robustness exhibited in human perception. Our experiments show that by leverag
ing sparsity and modeling the biological mechanisms at a cellular level, we are
able to mitigate the effect of adversarial alterations to the signal that have n
o perceptible meaning. Furthermore, we present and illustrate the effects of top
-down functional processes that contribute to the inherent immunity in human per
ception in the context of exploiting these properties to make a more robust mach

ine vision system.
****
Rethinking Differentiable Search for Mixed-Precision Neural Networks
Zhaowei Cai, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2349-2358
Low-precision networks, with weights and activations quantized to low bit-width, are widely used to accelerate inference on edge devices. However, current solutions are uniform, using identical bit-width for all filters. This fails to account for the different sensitivities of different filters and is suboptimal. Mixed-precision networks address this problem, by tuning the bit-width to individual filter requirements. In this work, the problem of optimal mixed-precision network search (MPS) is considered. To circumvent its difficulties of discrete search space and combinatorial optimization, a new differentiable search architecture is proposed, with several novel contributions to advance the efficiency by leveraging the unique properties of the MPS problem. The resulting Efficient differentiable MIxed-Precision network Search (EdMIPS) method is effective at finding the optimal bit allocation for multiple popular networks, and can search a large model, e.g. Inception-V3, directly on ImageNet without proxy task in a reasonable amount of time. The learned mixed-precision networks significantly outperform their uniform counterparts.
****
Wavelet Synthesis Net for Disparity Estimation to Synthesize DSLR Calibre Bokeh Effect on Smartphones
Chenchi Luo, Yingmao Li, Kaimo Lin, George Chen, Seok-Jun Lee, Jihwan Choi, Youngjun Francis Yoo, Michael O. Polley; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2407-2415
Modern smartphone cameras can match traditional DSLR cameras in many areas thanks to the introduction of camera arrays and multi-frame processing. Among all types of DSLR effects, the narrow depth of field (DoF) or so called bokeh probably arouses most interest. Today's smartphones try to overcome the physical lens and sensor limitations by introducing computational methods that utilize a depth map to synthesize the narrow DoF effect from all-in-focus images. However, a high quality depth map remains to be the key differentiator between computational bokeh and DSLR optical bokeh. Empowered by a novel wavelet synthesis network architecture, we have greatly narrowed the gap between DSLR and smartphone camera in terms of the bokeh more than ever before. We describe three key Modern smartphone cameras can match traditional digital single lens reflex (DSLR) cameras in many areas thanks to the introduction of camera arrays and multi-frame processing. Among all types of DSLR effects, the narrow depth of field (DoF) or so called bokeh probably arouses most interest. Today's smartphones try to overcome the physical lens and sensor limitations by introducing computational methods that utilize a depth map to synthesize the narrow DoF effect from all-in-focus images. However, a high quality depth map remains to be the key differentiator between computational bokeh and DSLR optical bokeh. Empowered by a novel wavelet synthesis network architecture, we have narrowed the gap between DSLR and smartphone camera in terms of bokeh more than ever before. We describe three key enablers of our bokeh solution: a synthetic graphics engine to generate training data with precisely prescribed characteristics that match the real smartphone captures, a novel wavelet synthesis neural network (WSN) architecture to produce unprecedented high definition disparity map promptly on smartphones, and a new evaluation metric to quantify the quality of the disparity map for real images from the bokeh rendering perspective. Experimental results show that the disparity map produced from our neural network achieves much better accuracy than the other state-of-the-art CNN based algorithms. Combining the high resolution disparity map with our rendering algorithm, we demonstrate visually superior bokeh pictures compared with existing top rated flagship smartphones listed on the DXOMARK mobiles.
****
Structure-Guided Ranking Loss for Single Image Depth Prediction
Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, Zhiguo Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP

R), 2020, pp. 611-620
Single image depth prediction is a challenging task due to its ill-posed nature and challenges with capturing ground truth for supervision. Large-scale disparity data generated from stereo photos and 3D videos is a promising source of supervision, however, such disparity data can only approximate the inverse ground truth depth up to an affine transformation. To more effectively learn from such pseudo-depth data, we propose to use a simple pair-wise ranking loss with a novel sampling strategy. Instead of randomly sampling point pairs, we guide the sampling to better characterize structure of important regions based on the low-level edge maps and high-level object instance masks. We show that the pair-wise ranking loss, combined with our structure-guided sampling strategies, can significantly improve the quality of depth map prediction. In addition, we introduce a new relative depth dataset of about 21K diverse high-resolution web stereo photos to enhance the generalization ability of our model. In experiments, we conduct cross-dataset evaluation on six benchmark datasets and show that our method consistently improves over the baselines, leading to superior quantitative and qualitative results.
****

Perspective Plane Program Induction From a Single Image
Yikai Li, Jiayuan Mao, Xiuming Zhang, William T. Freeman, Joshua B. Tenenbaum, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4434-4443
We study the inverse graphics problem of inferring a holistic representation for natural images. Given an input image, our goal is to induce a neuro-symbolic, program-like representation that jointly models camera poses, object locations, and global scene structures. Such high-level, holistic scene representations further facilitate low-level image manipulation tasks such as inpainting. We formulate this problem as jointly finding the camera pose and scene structure that best describe the input image. The benefits of such joint inference are two-fold: scene regularity serves as a new cue for perspective correction, and in turn, correct perspective correction leads to a simplified scene structure, similar to how the correct shape leads to the most regular texture in shape from texture. Our proposed framework, Perspective Plane Program Induction (P3I), combines search-based and gradient-based algorithms to efficiently solve the problem. P3I outperforms a set of baselines on a collection of Internet images, across tasks including camera pose estimation, global structure inference, and down-stream image manipulation tasks.
****

ActionBytes: Learning From Trimmed Videos to Localize Actions
Mihir Jain, Amir Ghodrati, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1171-1180
This paper tackles the problem of localizing actions in long untrimmed videos. Different from existing works, which all use annotated untrimmed videos during training, we learn only from short trimmed videos. This enables learning from large-scale datasets originally designed for action classification. We propose a method to train an action localization network that segments a video into interpretable fragments, we call ActionBytes. Our method jointly learns to cluster ActionBytes and trains the localization network using the cluster assignments as pseudo-labels. By doing so, we train on short trimmed videos that become untrimmed for ActionBytes. In isolation, or when merged, the ActionBytes also serve as effective action proposals. Experiments demonstrate that our boundary-guided training generalizes to unknown action classes and localizes actions in long videos of Thumos14, MultiThumos, and ActivityNet1.2. Furthermore, we show the advantage of ActionBytes for zero-shot localization as well as traditional weakly supervised localization, that train on long videos, to achieve state-of-the-art results.
****

Conv-MPN: Convolutional Message Passing Neural Network for Structured Outdoor Architecture Reconstruction
Fuyang Zhang, Nelson Nauata, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2798-2807

This paper proposes a novel message passing neural (MPN) architecture Conv-MPN, which reconstructs an outdoor building as a planar graph from a single RGB image . Conv-MPN is specifically designed for cases where nodes of a graph have explic it spatial embedding. In our problem, nodes correspond to building edges in an i mage. Conv-MPN is different from MPN in that 1) the feature associated with a no de is represented as a feature volume instead of a 1D vector; and 2) convolution s encode messages instead of fully connected layers. Conv-MPN learns to select a true subset of nodes (i.e., building edges) to reconstruct a building planar gr aph. Our qualitative and quantitative evaluations over 2,000 buildings show that Conv-MPN makes significant improvements over the existing fully neural solution s. We believe that the paper has a potential to open a new line of graph neural network research for structured geometry reconstruction.

****

Novel Object Viewpoint Estimation Through Reconstruction Alignment

Mohamed El Banani, Jason J. Corso, David F. Fouhey; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3113-3 122

The goal of this paper is to estimate the viewpoint for a novel object. Standard viewpoint estimation approaches generally fail on this task due to their relian ce on a 3D model for alignment or large amounts of class-specific training data and their corresponding canonical pose. We overcome those limitations by learnin g a reconstruct and align approach. Our key insight is that although we do not h ave an explicit 3D model or a predefined canonical pose, we can still learn to e stimate the object's shape in the viewer's frame and then use an image to provid e our reference model or canonical pose. In particular, we propose learning two networks: the first maps images to a 3D geometry-aware feature bottleneck and is trained via an image-to-image translation loss; the second learns whether two i nstances of features are aligned. At test time, our model finds the relative tra nsformation that best aligns the bottleneck features of our test image to a refe rence image. We evaluate our method on novel object viewpoint estimation by gene ralizing across different datasets, analyzing the impact of our different module s, and providing a qualitative analysis of the learned features to identify what representation is being learnt for alignment.

****

PaStaNet: Toward Human Activity Knowledge Engine

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Sh u Fang, Ze Ma, Mingyang Chen, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 382-391

Existing image-based activity understanding methods mainly adopt direct mapping, i.e. from image to activity concepts, which may encounter performance bottlenec k since the huge gap. In light of this, we propose a new path: infer human part states first and then reason out the activities based on part-level semantics. H uman Body Part States (PaSta) are fine-grained action semantic tokens, e.g. , wh ich can compose the activities and help us step toward human activity knowledge engine. To fully utilize the power of PaSta, we build a large-scale knowledge ba se PaStaNet, which contains 7M+ PaSta annotations. And two corresponding models are proposed: first, we design a model named Activity2Vec to extract PaSta featu res, which aim to be general representations for various activities. Second, we use a PaSta-based Reasoning method to infer activities. Promoted by PaStaNet, ou r method achieves significant improvements, e.g. 6.4 and 13.9 mAP on full and on e-shot sets of HICO in supervised learning, and 3.2 and 4.2 mAP on V-COCO and im ages-based AVA in transfer learning. Code and data are available at http://hake-mvig.cn/.

****

Dynamic Fluid Surface Reconstruction Using Deep Neural Network

Simron Thapa, Nianyi Li, Jinwei Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 21-30

Recovering the dynamic fluid surface is a long-standing challenging problem in c omputer vision. Most existing image-based methods require multiple views or a de dicated imaging system. Here we present a learning-based single-image approach f

or 3D fluid surface reconstruction. Specifically, we design a deep neural networ
k that estimates the depth and normal maps of a fluid surface by analyzing the r
efractive distortion of a reference background image. Due to the dynamic nature
of fluid surfaces, our network uses recurrent layers that carry temporal informa
tion from previous frames to achieve spatio-temporally consistent reconstruction
 given a video input. Due to the lack of fluid data, we synthesize a large fluid
 dataset using physics-based fluid modeling and rendering techniques for network
 training and validation. Through experiments on simulated and real captured flu
id images, we demonstrate that our proposed deep neural network trained on our f
luid dataset can recover dynamic 3D fluid surfaces with high accuracy.
****

MPM: Joint Representation of Motion and Position Map for Cell Tracking
Junya Hayashida, Kazuya Nishimura, Ryoma Bise; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3823-3832
Conventional cell tracking methods detect multiple cells in each frame (detectio
n) and then associate the detection results in successive time-frames (associati
on). Most cell tracking methods perform the association task independently from
the detection task. However, there is no guarantee of preserving coherence betwe
en these tasks, and lack of coherence may adversely affect tracking performance.
 In this paper, we propose the Motion and Position Map (MPM) that jointly repres
ents both detection and association for not only migration but also cell divisio
n. It guarantees coherence such that if a cell is detected, the corresponding mo
tion flow can always be obtained. It is a simple but powerful method for multi-o
bject tracking in dense environments. We compared the proposed method with curre
nt tracking methods under various conditions in real biological images and found
 that it outperformed the state-of-the-art (+5.2% improvement compared to the se
cond-best).
****

AdderNet: Do We Really Need Multiplications in Deep Learning?
Hanting Chen, Yunhe Wang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, Chang
Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2020, pp. 1468-1477
Compared with cheap addition operation, multiplication operation is of much high
er computation complexity. The widely-used convolutions in deep neural networks
are exactly cross-correlation to measure the similarity between input feature an
d convolution filters, which involves massive multiplications between float valu
es. In this paper, we present adder networks (AdderNets) to trade these massive
multiplications in deep neural networks, especially convolutional neural network
s (CNNs), for much cheaper additions to reduce computation costs. In AdderNets,
we take the L1-norm distance between filters and input feature as the output res
ponse. The influence of this new similarity measure on the optimization of neura
l network have been thoroughly analyzed. To achieve a better performance, we dev
elop a special back-propagation approach for AdderNets by investigating the full
-precision gradient. We then propose an adaptive learning rate strategy to enhan
ce the training procedure of AdderNets according to the magnitude of each neuron
's gradient. As a result, the proposed AdderNets can achieve 74.9% Top-1 accurac
y 91.7% Top-5 accuracy using ResNet-50 on the ImageNet dataset without any multi
plication in convolutional layer. The codes are publicly available at: (https://
github.com/huaweinoah/AdderNet).
****

Adaptive Interaction Modeling via Graph Operations Search
Haoxin Li, Wei-Shi Zheng, Yu Tao, Haifeng Hu, Jian-Huang Lai; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020
, pp. 521-530
Interaction modeling is important for video action analysis. Recently, several w
orks design specific structures to model interactions in videos. However, their
structures are manually designed and non-adaptive, which require structures desi
gn efforts and more importantly could not model interactions adaptively. In this
 paper, we automate the process of structures design to learn adaptive structure
s for interaction modeling. We propose to search the network structures with dif

ferentiable architecture search mechanism, which learns to construct adaptive st ructures for different videos to facilitate adaptive interaction modeling. To th is end, we first design the search space with several basic graph operations tha t explicitly capture different relations in videos. We experimentally demonstrat e that our architecture search framework learns to construct adaptive interactio n modeling structures, which provides more understanding about the relations bet ween the structures and some interaction characteristics, and also releases the requirement of structures design efforts. Additionally, we show that the designe d basic graph operations in the search space are able to model different interac tions in videos. The experiments on two interaction datasets show that our metho d achieves competitive performance with state-of-the-arts.
****

Differentiable Volumetric Rendering: Learning Implicit 3D Representations Withou t 3D Supervision
Michael Niemeyer, Lars Mescheder, Michael Oechsle, Andreas Geiger; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3504-3515
Learning-based 3D reconstruction methods have shown impressive results. However, most methods require 3D supervision which is often hard to obtain for real-worl d datasets. Recently, several works have proposed differentiable rendering techn iques to train reconstruction models from RGB images. Unfortunately, these appro aches are currently restricted to voxel- and mesh-based representations, sufferi ng from discretization or low resolution. In this work, we propose a differentia ble rendering formulation for implicit shape and texture representations. Implic it representations have recently gained popularity as they represent shape and t exture continuously. Our key insight is that depth gradients can be derived anal ytically using the concept of implicit differentiation. This allows us to learn implicit shape and texture representations directly from RGB images. We experime ntally show that our single-view reconstructions rival those learned with full 3 D supervision. Moreover, we find that our method can be used for multi-view 3D r econstruction, directly resulting in watertight meshes.
****

Memory-Efficient Hierarchical Neural Architecture Search for Image Denoising
Haokui Zhang, Ying Li, Hao Chen, Chunhua Shen; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3657-3666
Recently, neural architecture search (NAS) methods have attracted much attention and outperformed manually designed architectures on a few high-level vision tas ks. In this paper, we propose HiNAS (Hierarchical NAS), an effort towards employ ing NAS to automatically design effective neural network architectures for image denoising. HiNAS adopts gradient based search strategies and employs operations with adaptive receptive field to build an flexible hierarchical search space. D uring the search stage, HiNAS shares cells across different feature levels to sa ve memory and employ an early stopping strategy to avoid the collapse issue in N AS, and considerably accelerate the search speed. The proposed HiNAS is both mem ory and computation efficient, which takes only about 4.5 hours for searching us ing a single GPU. We evaluate the effectiveness of our proposed HiNAS on two dif ferent datasets, namely an additive white Gaussian noise dataset BSD500, and a r ealistic noise datasetSIM1800. Experimental results show that the architecture f ound by HiNAS has fewer parameters and enjoys a faster inference speed, while ac hieving highly competitive performance compared with state-of-the-art methods. W e also present analysis on the architectures found by NAS. HiNAS also shows good performance on experiments for image de-raining.
****

Embodied Language Grounding With 3D Visual Feature Representations
Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Ad am W. Harley, Katerina Fragkiadaki; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2020, pp. 2220-2229
We propose associating language utterances to 3D visual abstractions of the scen e they describe. The 3D visual abstractions are encoded as 3-dimensional visual feature maps. We infer these 3D visual scene feature maps from RGB images of the

scene via view prediction: when the generated 3D scene feature map is neurally projected from a camera viewpoint, it should match the corresponding RGB image. We present generative models that condition on the dependency tree of an utterance and generate a corresponding visual 3D feature map as well as reason about its plausibility, and detector models that condition on both the dependency tree of an utterance and a related image and localize the object referents in the 3D feature map inferred from the image. Our model outperforms models of language and vision that associate language with 2D CNN activations or 2D images by a large margin in a variety of tasks, such as, classifying plausibility of utterances, detecting referential expressions, and supplying rewards for trajectory optimization of object placement policies from language instructions. We perform numerous ablations and show the improved performance of our detectors is due to its better generalization across camera viewpoints and lack of object interferences in the inferred 3D feature space, and the improved performance of our generators is due to their ability to spatially reason about objects and their configurations in 3D when mapping from language to scenes.
****

Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching
Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2495-2504
The deep multi-view stereo (MVS) and stereo matching approaches generally construct 3D cost volumes to regularize and regress the output depth or disparity. These methods are limited when high-resolution outputs are needed since the memory and time costs grow cubically as the volume resolution increases. In this paper, we propose a both memory and time efficient cost volume formulation that is complementary to existing multi-view stereo and stereo matching approaches based on 3D cost volumes. First, the proposed cost volume is built upon a standard feature pyramid encoding geometry and context at gradually finer scales. Then, we can narrow the depth (or disparity) range of each stage by the depth (or disparity) map from the previous stage. With gradually higher cost volume resolution and adaptive adjustment of depth (or disparity) intervals, the output is recovered in a coarser to fine manner. We apply the cascade cost volume to the representative MVS-Net, and obtain a 35.6% improvement on DTU benchmark (1st place), with 50.6% and 59.3% reduction in GPU memory and run-time. It is also the state-of-the-art learning-based method on Tanks and Temples benchmark. The statistics of accuracy, run-time and GPU memory on other representative stereo CNNs also validate the effectiveness of our proposed method. Our source code is available at https://github.com/alibaba/cascade-stereo.
****

All in One Bad Weather Removal Using Architectural Search
Ruoteng Li, Robby T. Tan, Loong-Fah Cheong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3175-3185
Many methods have set state-of-the-art performance on restoring images degraded by bad weather such as rain, haze, fog, and snow, however they are designed specifically to handle one type of degradation. In this paper, we propose a method that can handle multiple bad weather degradations: rain, fog, snow and adherent raindrops using a single network. To achieve this, we first design a generator with multiple task-specific encoders, each of which is associated with a particular bad weather degradation type. We utilize a neural architecture search to optimally process the image features extracted from all encoders. Subsequently, to convert degraded image features to clean background features, we introduce a series of tensor-based operations encapsulating the underlying physics principles behind the formation of rain, fog, snow and adherent raindrops. These operations serve as the basic building blocks for our architectural search. Finally, our discriminator simultaneously assesses the correctness and classifies the degradation type of the restored image. We design a novel adversarial learning scheme that only backpropagates the loss of a degradation type to the respective task-specific encoder. Despite being designed to handle different types of bad weather, extensive experiments demonstrate that our method performs competitively to the ind

ividual and dedicated state-of-the-art image restoration methods.
****
Fast-MVSNet: Sparse-to-Dense Multi-View Stereo With Learned Propagation and Gauss-Newton Refinement
Zehao Yu, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1949-1958
Almost all previous deep learning-based multi-view stereo (MVS) approaches focus on improving reconstruction quality. Besides quality, efficiency is also a desirable feature for MVS in real scenarios. Towards this end, this paper presents a Fast-MVSNet, a novel sparse-to-dense coarse-to-fine framework, for fast and accurate depth estimation in MVS. Specifically, in our Fast-MVSNet, we first construct a sparse cost volume for learning a sparse and high-resolution depth map. Then we leverage a small-scale convolutional neural network to encode the depth dependencies for pixels within a local region to densify the sparse high-resolution depth map. At last, a simple but efficient Gauss-Newton layer is proposed to further optimize the depth map. On one hand, the high-resolution depth map, the data-adaptive propagation method and the Gauss-Newton layer jointly guarantee the effectiveness of our method. On the other hand, all modules in our Fast-MVSNet are lightweight and thus guarantee the efficiency of our approach. Besides, our approach is also memory-friendly because of the sparse depth representation. Extensive experimental results show that our method is 5 times and 14 times faster than Point-MVSNet and R-MVSNet, respectively, while achieving comparable or even better results on the challenging Tanks and Temples dataset as well as the DTU dataset. Code is available at https://github.com/svip-lab/FastMVSNet.
****
Auxiliary Training: Towards Accurate and Robust Models
Linfeng Zhang, Muzhou Yu, Tong Chen, Zuoqiang Shi, Chenglong Bao, Kaisheng Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 372-381
Training process is crucial for the deployment of the network in applications which have two strict requirements on both accuracy and robustness. However, most existing approaches are in a dilemma, i.e. model accuracy and robustness form an embarrassing tradeoff - the improvement of one leads to the drop of the other. The challenge remains as for we try to improve the accuracy and robustness simultaneously. In this paper, we propose a novel training method via introducing the auxiliary classifiers for training on corrupted samples, while the clean samples are normally trained with the primary classifier. In the training stage, a novel distillation method named input-aware self distillation is proposed to facilitate the primary classifier to learn the robust information from auxiliary classifiers. Along with it, a new normalization method - selective batch normalization is proposed to prevent the model from the negative influence of corrupted images. At the end of training period, a L2-norm penalty is applied to the weights of primary and auxiliary classifiers such that their weights are asymptotically identical. In the stage of inference, only the primary classifier is used and thus no extra computation and storage are needed. Extensive experiments on CIFAR10, CIFAR100 and ImageNet show that noticeable improvements on both accuracy and robustness can be observed by the proposed auxiliary training. On average, auxiliary training achieves 2.21% accuracy and 21.64% robustness (measured by corruption error) improvements over traditional training methods on CIFAR100. Codes has been released on github.
****
Cascaded Human-Object Interaction Recognition
Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4263-4272
Rapid progress has been witnessed for human-object interaction (HOI) recognition, but most existing models are confined to single-stage reasoning pipelines. Considering the intrinsic complexity of the task, we introduce a cascade architecture for a multi-stage, coarse-to-fine HOI understanding. At each stage, an instance localization network progressively refines HOI proposals and feeds them into

an interaction recognition network. Each of the two networks is also connected t
o its predecessor at the previous stage, enabling cross-stage information propag
ation. The interaction recognition network has two crucial parts: a relation ran
king module for high-quality HOI proposal selection and a triple-stream classifi
er for relation prediction. With our carefully-designed human-centric relation f
eatures, these two modules work collaboratively towards effective interaction un
derstanding. Further beyond relation detection on a bounding-box level, we make
our framework flexible to perform fine-grained pixel-wise relation segmentation;
 this provides a new glimpse into better relation modeling. Our approach reached
 the 1^ st place in the ICCV2019 Person in Context Challenge, on both relation
detection and segmentation tasks. It also shows promising results on V-COCO.
****

Holistically-Attracted Wireframe Parsing
Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, P
hilip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2020, pp. 2788-2797
This paper presents a fast and parsimonious parsing method to accurately and rob
ustly detect a vectorized wireframe in an input image with a single forward pass
. The proposed method is end-to-end trainable, consisting of three components: (
i) line segment and junction proposal generation, (ii) line segment and junction
 matching, and (iii) line segment and junction verification. For computing line
segment proposals, a novel exact dual representation is proposed which exploits
a parsimonious geometric reparameterization for line segments and forms a holist
ic 4-dimensional attraction field map for an input image. Junctions can be treat
ed as the "basins" in the attraction field. The proposed method is thus called H
olistically-Attracted Wireframe Parser (HAWP). In experiments, the proposed meth
od is tested on two benchmarks, the Wireframe dataset [14] and the YorkUrban dat
aset [8]. On both benchmarks, it obtains state-of-the-art performance in terms o
f accuracy and efficiency. For example, on the Wireframe dataset, compared to th
e previous state-of-the-art method L-CNN [36], it improves the challenging mean
structural average precision (msAP) by a large margin (2.8% absolute improvement
s), and achieves 29.5 FPS on a single GPU (89% relative improvement). A systemat
ic ablation study is performed to further justify the proposed method.
****

Strip Pooling: Rethinking Spatial Pooling for Scene Parsing
Qibin Hou, Li Zhang, Ming-Ming Cheng, Jiashi Feng; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4003-4
012
Spatial pooling has been proven highly effective to capture long-range contextua
l information for pixel-wise prediction tasks, such as scene parsing. In this pa
per, beyond conventional spatial pooling that usually has a regular shape of NxN
, we rethink the formulation of spatial pooling by introducing a new pooling str
ategy, called strip pooling, which considers a long but narrow kernel, i.e., 1xN
 or Nx1. Based on strip pooling, we further investigate spatial pooling architec
ture design by 1) introducing a new strip pooling module that enables backbone n
etworks to efficiently model long-range dependencies; 2) presenting a novel buil
ding block with diverse spatial pooling as a core; and 3) systematically compari
ng the performance of the proposed strip pooling and conventional spatial poolin
g techniques. Both novel pooling-based designs are lightweight and can serve as
an efficient plug-and-play modules in existing scene parsing networks. Extensive
 experiments on Cityscapes and ADE20K benchmarks demonstrate that our simple app
roach establishes new state-of-the-art results. Code is available at https://git
hub.com/Andrew-Qibin/SPNet.
****

OASIS: A Large-Scale Dataset for Single Image 3D in the Wild
Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, Jia D
eng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2020, pp. 679-688
Single-view 3D is the task of recovering 3D properties such as depth and surface
 normals from a single image. We hypothesize that a major obstacle to single-ima

ge 3D is data. We address this issue by presenting Open Annotations of Single Image Surfaces (OASIS), a dataset for single-image 3D in the wild consisting of annotations of detailed 3D geometry for 140,000 images. We train and evaluate leading models on a variety of single-image 3D tasks. We expect OASIS to be a useful resource for 3D vision research. Project site: https://pvl.cs.princeton.edu/OASIS.

****

CONSAC: Robust Multi-Model Fitting by Conditional Sample Consensus

Florian Kluger, Eric Brachmann, Hanno Ackermann, Carsten Rother, Michael Ying Yang, Bodo Rosenhahn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4634-4643

We present a robust estimator for fitting multiple parametric models of the same form to noisy measurements. Applications include finding multiple vanishing points in man-made scenes, fitting planes to architectural imagery, or estimating multiple rigid motions within the same sequence. In contrast to previous works, which resorted to hand-crafted search strategies for multiple model detection, we learn the search strategy from data. A neural network conditioned on previously detected models guides a RANSAC estimator to different subsets of all measurements, thereby finding model instances one after another. We train our method supervised, as well as, self-supervised. For supervised training of the search strategy, we contribute a new dataset for vanishing point estimation. Leveraging this dataset, the proposed algorithm is superior with respect to other robust estimators, as well as, to designated vanishing point estimation algorithms. For self-supervised learning of the search, we evaluate the proposed algorithm on multi-homography estimation and demonstrate an accuracy that is superior to state-of-the-art methods.

****

Deep Global Registration

Christopher Choy, Wei Dong, Vladlen Koltun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2514-2523

We present Deep Global Registration, a differentiable framework for pairwise registration of real-world 3D scans. Deep global registration is based on three modules: a 6-dimensional convolutional network for correspondence confidence prediction, a differentiable Weighted Procrustes algorithm for closed-form pose estimation, and a robust gradient-based SE(3) optimizer for pose refinement. Experiments demonstrate that our approach outperforms state-of-the-art methods, both learning-based and classical, on real-world data.

****

CycleISP: Real Image Restoration via Improved Data Synthesis

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), 2020, pp. 2696-2705

The availability of large-scale datasets has helped unleash the true potential of deep convolutional neural networks (CNNs). However, for the single-image denoising problem, capturing a real dataset is an unacceptably expensive and cumbersome procedure. Consequently, image denoising algorithms are mostly developed and evaluated on synthetic data that is usually generated with a widespread assumption of additive white Gaussian noise (AWGN). While the CNNs achieve impressive results on these synthetic datasets, they do not perform well when applied on real camera images, as reported in recent benchmark datasets. This is mainly because the AWGN is not adequate for modeling the real camera noise which is signal-dependent and heavily transformed by the camera imaging pipeline. In this paper, we present a framework that models camera imaging pipeline in forward and reverse directions. It allows us to produce any number of realistic image pairs for denoising both in RAW and sRGB spaces. By training a new image denoising network on realistic synthetic data, we achieve the state-of-the-art performance on real camera benchmark datasets. The parameters in our models are 5 times lesser than the previous best method for RAW denoising. Furthermore, we demonstrate that the proposed framework generalizes beyond image denoising problem e.g., for color matching in stereoscopic cinema. The source code and pre-trained models are availa

ble at https://github.com/swz30/CycleISP.
****
Neural Network Pruning With Residual-Connections and Limited-Data
Jian-Hao Luo, Jianxin Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1458-1467
Filter level pruning is an effective method to accelerate the inference speed of deep CNN models. Although numerous pruning algorithms have been proposed, there are still two open issues. The first problem is how to prune residual connections. We propose to prune both channels inside and outside the residual connections via a KL-divergence based criterion. The second issue is pruning with limited data. We observe an interesting phenomenon: directly pruning on a small dataset is usually worse than fine-tuning a small model which is pruned or trained from scratch on the large dataset. Knowledge distillation is an effective approach to compensate for the weakness of limited data. However, the logits of a teacher model may be noisy. In order to avoid the influence of label noise, we propose a label refinement approach to solve this problem. Experiments have demonstrated the effectiveness of our method (CURL, Compression Using Residual-connections and Limited-data). CURL significantly outperforms previous state-of-the-art methods on ImageNet. More importantly, when pruning on small datasets, CURL achieves comparable or much better performance than fine-tuning a pretrained small model.
****
Neural Cages for Detail-Preserving 3D Deformations
Wang Yifan, Noam Aigerman, Vladimir G. Kim, Siddhartha Chaudhuri, Olga Sorkine-Hornung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 75-83
We propose a novel learnable representation for detail preserving shape deformation. The goal of our method is to warp a source shape to match the general structure of a target shape, while preserving the surface details of the source. Our method extends a traditional cage-based deformation technique, where the source shape is enclosed by a coarse control mesh termed cage, and translations prescribed on the cage vertices are interpolated to any point on the source mesh via special weight functions. The use of this sparse cage scaffolding enables preserving surface details regardless of the shape's intricacy and topology. Our key contribution is a novel neural network architecture for predicting deformations by controlling the cage. We incorporate a differentiable cage-based deformation module in our architecture, and train our network end-to-end. Our method can be trained with common collections of 3D models in an unsupervised fashion, without any cage-specific annotations. We demonstrate the utility of our method for synthesizing shape variations and deformation transfer.
****
Fashion Outfit Complementary Item Retrieval
Yen-Liang Lin, Son Tran, Larry S. Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3311-3319
Complementary fashion item recommendation is critical for fashion outfit completion. Existing methods mainly focus on outfit compatibility prediction but not in a retrieval setting. We propose a new framework for outfit complementary item retrieval. Specifically, a category-based subspace attention network is presented, which is a scalable approach for learning the subspace attentions. In addition, we introduce an outfit ranking loss that better models the item relationships of an entire outfit. We evaluate our method on the outfit compatibility, FITB and new retrieval tasks. Experimental results demonstrate that our approach outperforms state-of-the-art methods in both compatibility prediction and complementary item retrieval.
****
SCT: Set Constrained Temporal Transformer for Set Supervised Action Segmentation
Mohsen Fayyaz, Jurgen Gall; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 501-510
Temporal action segmentation is a topic of increasing interest, however, annotating each frame in a video is cumbersome and costly. Weakly supervised approaches therefore aim at learning temporal action segmentation from videos that are onl

y weakly labeled. In this work, we assume that for each training video only the list of actions is given that occur in the video, but not when, how often, and in which order they occur. In order to address this task, we propose an approach that can be trained end-to-end on such data. The approach divides the video into smaller temporal regions and predicts for each region the action label and its length. In addition, the network estimates the action labels for each frame. By measuring how consistent the frame-wise predictions are with respect to the temporal regions and the annotated action labels, the network learns to divide a video into class-consistent regions. We evaluate our approach on three datasets where the approach achieves state-of-the-art results.
****

CPR-GCN: Conditional Partial-Residual Graph Convolutional Network in Automated Anatomical Labeling of Coronary Arteries
Han Yang, Xingjian Zhen, Ying Chi, Lei Zhang, Xian-Sheng Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3803-3811
Automated anatomical labeling plays a vital role in coronary artery disease diagnosing procedure. The main challenge in this problem is the large individual variability inherited in human anatomy. Existing methods usually rely on the position information and the prior knowledge of the topology of the coronary artery tree, which may lead to unsatisfactory performance when the main branches are confusing. Motivated by the wide application of the graph neural network in structured data, in this paper, we propose a conditional partial-residual graph convolutional network (CPR-GCN), which takes both position and CT image into consideration, since CT image contains abundant information such as branch size and spanning direction. Two majority parts, a Partial-Residual GCN and a conditions extractor, are included in CPR-GCN. The conditions extractor is a hybrid model containing the 3D CNN and the LSTM, which can extract 3D spatial image features along the branches. On the technical side, the Partial-Residual GCN takes the position features of the branches, with the 3D spatial image features as conditions, to predict the label for each branches. While on the mathematical side, our approach twists the partial differential equation (PDE) into the graph modeling. A dataset with 511 subjects is collected from the clinic and annotated by two experts with a two-phase annotation process. According to the five-fold cross-validation, our CPR-GCN yields 95.8% meanRecall, 95.4% meanPrecision and 0.955 meanF1, which outperforms state-of-the-art approaches.
****

Neural Blind Deconvolution Using Deep Priors
Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3341-3350
Blind deconvolution is a classical yet challenging low-level vision problem with many real-world applications. Traditional maximum a posterior (MAP) based methods rely heavily on fixed and handcrafted priors that certainly are insufficient in characterizing clean images and blur kernels, and usually adopt specially designed alternating minimization to avoid trivial solution. In contrast, existing deep motion deblurring networks learn from massive training images the mapping to clean image or blur kernel, but are limited in handling various complex and large size blur kernels. To connect MAP and deep models, we in this paper present two generative networks for respectively modeling the deep priors of clean image and blur kernel, and propose an unconstrained neural optimization solution to blind deconvolution. In particular, we adopt an asymmetric Autoencoder with skip connections for generating latent clean image, and a fully-connected network (FCN) for generating blur kernel. Moreover, the SoftMax nonlinearity is applied to the output layer of FCN to meet the non-negative and equality constraints. The process of neural optimization can be explained as a kind of "zero-shot" self-supervised learning of the generative networks, and thus our proposed method is dubbed SelfDeblur. Experimental results show that our SelfDeblur can achieve notable quantitative gains as well as more visually plausible deblurring results in comparison to state-of-the-art blind deconvolution methods on benchmark datasets a

nd real-world blurry images. The source code is publicly available at https://github.com/csdwren/SelfDeblur
****

3D Packing for Self-Supervised Monocular Depth Estimation
Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2485-2494
Although cameras are ubiquitous, robotic platforms typically rely on active sensors like LiDAR for direct 3D perception. In this work, we propose a novel self-supervised monocular depth estimation method combining geometry with a new deep network, PackNet, learned only from unlabeled monocular videos. Our architecture leverages novel symmetrical packing and unpacking blocks to jointly learn to compress and decompress detail-preserving representations using 3D convolutions. Although self-supervised, our method outperforms other self, semi, and fully supervised methods on the KITTI benchmark. The 3D inductive bias in PackNet enables it to scale with input resolution and number of parameters without overfitting, generalizing better on out-of-domain data such as the NuScenes dataset. Furthermore, it does not require large-scale supervised pretraining on ImageNet and can run in real-time. Finally, we release DDAD (Dense Depth for Automated Driving), a new urban driving dataset with more challenging and accurate depth evaluation, thanks to longer-range and denser ground-truth depth generated from high-density LiDARs mounted on a fleet of self-driving cars operating world-wide.
****

Local Non-Rigid Structure-From-Motion From Diffeomorphic Mappings
Shaifali Parashar, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2059-2067
We propose a new formulation to non-rigid structure-from-motion that only requires the deforming surface to preserve its differential structure. This is a much weaker assumption than the traditional ones of isometry or conformality. We show that it is nevertheless sufficient to establish local correspondences between the surface in two different images and therefore to perform point-wise reconstruction using only first-order derivatives. To this end, we formulate differential constraints and solve them algebraically using the theory of resultants. We will demonstrate that our approach is more widely applicable, more stable in noisy and sparse imaging conditions and much faster than earlier ones, while delivering similar accuracy. The code is available at https://github.com/cvlab-epfl/diff-nrsfm/.
****

Structure Preserving Generative Cross-Domain Learning
Haifeng Xia, Zhengming Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4364-4373
Unsupervised domain adaptation (UDA) casts a light when dealing with insufficient or no labeled data in the target domain by exploring the well-annotated source knowledge in different distributions. Most research efforts on UDA explore to seek a domain-invariant classifier over source supervision. However, due to the scarcity of label information in the target domain, such a classifier has a lack of ground-truth target supervision, which dramatically obstructs the robustness and discrimination of the classifier. To this end, we develop a novel Generative cross-domain learning via Structure-Preserving (GSP), which attempts to transform target data into the source domain in order to take advantage of source supervision. Specifically, a novel cross-domain graph alignment is developed to capture the intrinsic relationship across two domains during target-source translation. Simultaneously, two distinct classifiers are trained to trigger the domain-invariant feature learning both guided with source supervision, one is a traditional source classifier and the other is a source-supervised target classifier. Extensive experimental results on several cross-domain visual benchmarks have demonstrated the effectiveness of our model by comparing with other state-of-the-art UDA algorithms.
****

Generative Hybrid Representations for Activity Forecasting With No-Regret Learni

ng
Jiaqi Guan, Ye Yuan, Kris M. Kitani, Nicholas Rhinehart; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 173-182
Automatically reasoning about future human behaviors is a difficult problem but has significant practical applications to assistive systems. Part of this difficulty stems from learning systems' inability to represent all kinds of behaviors. Some behaviors, such as motion, are best described with continuous representations, whereas others, such as picking up a cup, are best described with discrete representations. Furthermore, human behavior is generally not fixed: people can change their habits and routines. This suggests these systems must be able to learn and adapt continuously. In this work, we develop an efficient deep generative model to jointly forecast a person's future discrete actions and continuous motions. On a large-scale egocentric dataset, EPIC-KITCHENS, we observe our method generates high-quality and diverse samples while exhibiting better generalization than related generative models. Finally, we propose a variant to continually learn our model from streaming data, observe its practical effectiveness, and theoretically justify its learning efficiency.
****
Predicting Cognitive Declines Using Longitudinally Enriched Representations for Imaging Biomarkers
Lyujian Lu, Hua Wang, Saad Elbeleidy, Feiping Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4827-4836
With rapid progress in high-throughput genotyping and neuroimaging, researches of complex brain disorders, such as Alzheimer's Disease (AD), have gained significant attention in recent years. Many prediction models have been studied to relate neuroimaging measures to cognitive status over the progressions when these disease develops. Missing data is one of the biggest challenge in accurate cognitive score prediction of subjects in longitudinal neuroimaging studies. To tackle this problem, in this paper we propose a novel formulation to learn an enriched representation for imaging biomarkers that can simultaneously capture both the information conveyed by baseline neuroimaging records and that by progressive variations of varied counts of available follow-up records over time. While the numbers of the brain scans of the participants vary, the learned biomarker representation for every participant is a fixed-length vector, which enable us to use traditional learning models to study AD developments. Our new objective is formulated to maximize the ratio of the summations of a number of L1-norm distances for improved robustness, which, though, is difficult to efficiently solve in general. Thus we derive a new efficient iterative solution algorithm and rigorously prove its convergence. We have performed extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. A performance gain has been achieved to predict four different cognitive scores, when we compare the original baseline representations against the learned representations with enrichments. These promising empirical results have demonstrated improved performances of our new method that validate its effectiveness.
****
3D Sketch-Aware Semantic Scene Completion via Semi-Supervised Structure Prior
Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4193-4202
The goal of the Semantic Scene Completion (SSC) task is to simultaneously predict a completed 3D voxel representation of volumetric occupancy and semantic labels of objects in the scene from a single-view observation. Since the computational cost generally increases explosively along with the growth of voxel resolution, most current state-of-the-arts have to tailor their framework into a low-resolution representation with the sacrifice of detail prediction. Thus, voxel resolution becomes one of the crucial difficulties that lead to the performance bottleneck. In this paper, we propose to devise a new geometry-based strategy to embed depth information with low-resolution voxel representation, which could still b

e able to encode sufficient geometric information, e.g., room layout, object's s
izes and shapes, to infer the invisible areas of the scene with well structure-p
reserving details. To this end, we first propose a novel 3D sketch-aware feature
 embedding to explicitly encode geometric information effectively and efficientl
y. With the 3D sketch in hand, we further devise a simple yet effective semantic
 scene completion framework that incorporates a light-weight 3D Sketch Hallucina
tion module to guide the inference of occupancy and the semantic labels via a se
mi-supervised structure prior learning strategy. We demonstrate that our propose
d geometric embedding works better than the depth feature learning from habitual
 SSC frameworks. Our final model surpasses state- of-the-arts consistently on th
ree public benchmarks, which only requires 3D volumes of 60 x 36 x 60 resolution
 for both input and output.
****

Progressive Mirror Detection

Jiaying Lin, Guodong Wang, Rynson W.H. Lau; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3697-3705

The mirror detection problem is important as mirrors can affect the performances
 of many vision tasks. It is a difficult problem as it requires an understanding
 of global scene semantics. Recently, a method was proposed to detect mirrors by
 learning multi-level contextual contrasts between inside and outside of mirrors
, which helps locate mirror edges implicitly. We observe that the content of a m
irror reflects the content of its surrounding, separated by the edge of the mirr
or. Hence, we propose a model in this paper to progressively learn the content s
imilarity between the inside and outside of the mirror while explicitly detectin
g the mirror edges. Our work has two main contributions. First, we propose a new
 relational contextual contrasted local (RCCL) module to extract and compare the
 mirror features with its corresponding context features, and an edge detection
and fusion (EDF) module to learn the features of mirror edges in complex scenes
via explicit supervision. Second, we construct a challenging benchmark dataset o
f 6,461 mirror images. Unlike the existing MSD dataset, which has limited divers
ity, our dataset covers a variety of scenes and is much larger in scale. Experim
ental results show that our model outperforms relevant state-of-the-art methods.
****

FOAL: Fast Online Adaptive Learning for Cardiac Motion Estimation

Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas S. Huan
g, Terrence Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2020, pp. 4313-4323

Motion estimation of cardiac MRI videos is crucial for the evaluation of human h
eart anatomy and function. Recent researches show promising results with deep le
arning-based methods. In clinical deployment, however, they suffer dramatic perf
ormance drops due to mismatched distributions between training and testing datas
ets, commonly encountered in the clinical environment. On the other hand, it is
arguably impossible to collect all representative datasets and to train a univer
sal tracker before deployment. In this context, we proposed a novel fast online
adaptive learning (FOAL) framework: an online gradient descent based optimizer t
hat is optimized by a meta-learner. The meta-learner enables the online optimize
r to perform a fast and robust adaptation. We evaluated our method through exten
sive experiments on two public clinical datasets. The results showed the superio
r performance of FOAL in accuracy compared to the offline-trained tracking metho
d. On average, the FOAL took only 0.4 second per video for online optimization.
****

Don't Hit Me! Glass Detection in Real-World Scenes

Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang,
 Xiaopeng Wei, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2020, pp. 3687-3696

Glass is very common in our daily life. Existing computer vision systems neglect
 it and thus may have severe consequences, e.g., a robot may crash into a glass
wall. However, sensing the presence of glass is not straightforward. The key cha
llenge is that arbitrary objects/scenes can appear behind the glass, and the con
tent within the glass region is typically similar to those behind it. In this pa

per, we propose an important problem of detecting glass from a single RGB image. To address this problem, we construct a large-scale glass detection dataset (GDD) and design a glass detection network, called GDNet, which explores abundant contextual cues for robust glass detection with a novel large-field contextual feature integration (LCFI) module. Extensive experiments demonstrate that the proposed method achieves more superior glass detection results on our GDD test set than state-of-the-art methods fine-tuned for glass detection.
****
Deep Grouping Model for Unified Perceptual Parsing
Zhiheng Li, Wenxuan Bao, Jiayang Zheng, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4053-4063
The perceptual-based grouping process produces a hierarchical and compositional image representation that helps both human and machine vision systems recognize heterogeneous visual concepts. Examples can be found in the classical hierarchical superpixel segmentation or image parsing works. However, the grouping process is largely overlooked in modern CNN-based image segmentation networks due to many challenges, including the inherent incompatibility between the grid-shaped CNN feature map and the irregular-shaped perceptual grouping hierarchy. Overcoming these challenges, we propose a deep grouping model (DGM) that tightly marries the two types of representations and defines a bottom-up and a top-down process for feature exchanging. When evaluating the model on the recent Broden+ dataset for the unified perceptual parsing task, it achieves state-of-the-art results while having a small computational overhead compared to other contextual-based segmentation models. Furthermore, the DGM has better interpretability compared with modern CNN methods.
****
Inter-Task Association Critic for Cross-Resolution Person Re-Identification
Zhiyi Cheng, Qi Dong, Shaogang Gong, Xiatian Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2605-2615
Person images captured by unconstrained surveillance cameras often have low resolutions (LR). This causes the resolution mismatch problem when matched against the high-resolution (HR) gallery images, negatively affecting the performance of person re-identification (re-id). An effective approach is to leverage image super-resolution (SR) along with person re-id in a joint learning manner. However, this scheme is limited due to dramatically more difficult gradients backpropagation during training. In this paper, we introduce a novel model training regularisation method, called Inter-Task Association Critic (INTACT), to address this fundamental problem. Specifically, INTACT discovers the underlying association knowledge between image SR and person re-id, and leverages it as an extra learning constraint for enhancing the compatibility of SR model with person re-id in HR image space. This is realised by parameterising the association constraint which enables it to be automatically learned from the training data. Extensive experiments validate the superiority of INTACT over the state-of-the-art approaches on the cross-resolution re-id task using five standard person re-id datasets.
****
SOS: Selective Objective Switch for Rapid Immunofluorescence Whole Slide Image Classification
Sam Maksoud, Kun Zhao, Peter Hobson, Anthony Jennings, Brian C. Lovell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3862-3871
The difficulty of processing gigapixel whole slide images (WSIs) in clinical microscopy has been a long-standing barrier to implementing computer aided diagnostic systems. Since modern computing resources are unable to perform computations at this extremely large scale, current state of the art methods utilize patch-based processing to preserve the resolution of WSIs. However, these methods are often resource intensive and make significant compromises on processing time. In this paper, we demonstrate that conventional patch-based processing is redundant for certain WSI classification tasks where high resolution is only required in a

minority of cases. This reflects what is observed in clinical practice; where a pathologist may screen slides using a low power objective and only switch to a high power in cases where they are uncertain about their findings. To eliminate these redundancies, we propose a method for the selective use of high resolution processing based on the confidence of predictions on downscaled WSIs --- we call this the Selective Objective Switch (SOS). Our method is validated on a novel dataset of 684 Liver-Kidney-Stomach immunofluorescence WSIs routinely used in the investigation of autoimmune liver disease. By limiting high resolution processing to cases which cannot be classified confidently at low resolution, we maintain the accuracy of patch-level analysis whilst reducing the inference time by a factor of 7.74.

****

Learning Multi-Granular Hypergraphs for Video-Based Person Re-Identification

Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2899-2908

Video-based person re-identification (re-ID) is an important research topic in computer vision. The key to tackling the challenging task is to exploit both spatial and temporal clues in video sequences. In this work, we propose a novel graph-based framework, namely Multi-Granular Hypergraph (MGH), to pursue better representational capabilities by modeling spatiotemporal dependencies in terms of multiple granularities. Specifically, hypergraphs with different spatial granularities are constructed using various levels of part-based features across the video sequence. In each hypergraph, different temporal granularities are captured by hyperedges that connect a set of graph nodes (i.e., part-based features) across different temporal ranges. Two critical issues (misalignment and occlusion) are explicitly addressed by the proposed hypergraph propagation and feature aggregation schemes. Finally, we further enhance the overall video representation by learning more diversified graph-level representations of multiple granularities based on mutual information minimization. Extensive experiments on three widely-adopted benchmarks clearly demonstrate the effectiveness of the proposed framework. Notably, 90.0% top-1 accuracy on MARS is achieved using MGH, outperforming the state-of-the-arts.

****

Towards Unified INT8 Training for Convolutional Neural Network

Feng Zhu, Ruihao Gong, Fengwei Yu, Xianglong Liu, Yanfei Wang, Zhelong Li, Xiuqi Yang, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1969-1979

Recently low-bit (e.g., 8-bit) network quantization has been extensively studied to accelerate the inference. Besides inference, low-bit training with quantized gradients can further bring more considerable acceleration, since the backward process is often computation-intensive. Unfortunately, the inappropriate quantization of backward propagation usually makes the training unstable and even crash. There lacks a successful unified low-bit training framework that can support diverse networks on various tasks. In this paper, we give an attempt to build a unified 8-bit (INT8) training framework for common convolutional neural networks from the aspects of both accuracy and speed. First, we empirically find the four distinctive characteristics of gradients, which provide us insightful clues for gradient quantization. Then, we theoretically give an in-depth analysis of the convergence bound and derive two principles for stable INT8 training. Finally, we propose two universal techniques, including Direction Sensitive Gradient Clipping that reduces the direction deviation of gradients and Deviation Counteractive Learning Rate Scaling that avoids illegal gradient update along the wrong direction. The experiments show that our unified solution promises accurate and efficient INT8 training for a variety of networks and tasks, including MobileNetV2, InceptionV3 and object detection that prior studies have never succeeded. Moreover, it enjoys a strong flexibility to run on off-the-shelf hardware, and reduces the training time by 22% on Pascal GPU without too much optimization effort. We believe that this pioneering study will help lead the community towards a fully unified INT8 training for convolutional neural networks.

****
Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes
Sravanti Addepalli, Vivek B.S., Arya Baburaj, Gaurang Sriramanan, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1020-1029
As humans, we inherently perceive images based on their predominant features, and ignore noise embedded within lower bit planes. On the contrary, Deep Neural Networks are known to confidently misclassify images corrupted with meticulously crafted perturbations that are nearly imperceptible to the human eye. In this work, we attempt to address this problem by training networks to form coarse impressions based on the information in higher bit planes, and use the lower bit planes only to refine their prediction. We demonstrate that, by imposing consistency on the representations learned across differently quantized images, the adversarial robustness of networks improves significantly when compared to a normally trained model. Present state-of-the-art defenses against adversarial attacks require the networks to be explicitly trained using adversarial samples that are computationally expensive to generate. While such methods that use adversarial training continue to achieve the best results, this work paves the way towards achieving robustness without having to explicitly train on adversarial samples. The proposed approach is therefore faster, and also closer to the natural learning process in humans.
****
HOnnotate: A Method for 3D Annotation of Hand and Object Poses
Shreyas Hampali, Mahdi Rad, Markus Oberweger, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3196-3206
We propose a method for annotating images of a hand manipulating an object with the 3D poses of both the hand and the object, together with a dataset created using this method. Our motivation is the current lack of annotated real images for this problem, as estimating the 3D poses is challenging, mostly because of the mutual occlusions between the hand and the object. To tackle this challenge, we capture sequences with one or several RGB-D cameras and jointly optimize the 3D hand and object poses over all the frames simultaneously. This method allows us to automatically annotate each frame with accurate estimates of the poses, despite large mutual occlusions. With this method, we created HO-3D, the first markerless dataset of color images with 3D annotations for both the hand and object. This dataset is currently made of 77,558 frames, 68 sequences, 10 persons, and 10 objects. Using our dataset, we develop a single RGB image-based method to predict the hand pose when interacting with objects under severe occlusions and show it generalizes to objects not seen in the dataset.
****
From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality
Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, Alan Bovik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3575-3585
Blind or no-reference (NR) perceptual picture quality prediction is a difficult, unsolved problem of great consequence to the social and streaming media industries that impacts billions of viewers daily. Unfortunately, popular NR prediction models perform poorly on real-world distorted pictures. To advance progress on this problem, we introduce the largest (by far) subjective picture quality database, containing about 40, 000 real-world distorted pictures and 120, 000 patches, on which we collected about 4M human judgments of picture quality. Using these picture and patch quality labels, we built deep region-based architectures that learn to produce state-of-the-art global picture quality predictions as well as useful local picture quality maps. Our innovations include picture quality prediction architectures that produce global-to-local inferences as well as local-to-global inferences (via feedback). The dataset and source code are available at https: //live.ece.utexas.edu/research.php.

****
Fast MSER
Hailiang Xu, Siqi Xie, Fan Chen; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2020, pp. 3380-3389
Maximally Stable Extremal Regions (MSER) algorithms are based on the component t ree and are used to detect invariant regions. OpenCV MSER, the most popular MSER implementation, uses a linked list to associate pixels with ERs. The data-struc ture of an ER contains the attributes of a head and a tail linked node, which ma kes OpenCV MSER hard to be performed in parallel using existing parallel compone nt tree strategies. Besides, pixel extraction (i.e. extracting the pixels in MSE Rs) in OpenCV MSER is very slow. In this paper, we propose two novel MSER algori thms, called Fast MSER V1 and V2. They first divide an image into several spatia l partitions, then construct sub-trees and doubly linked lists (for V1) or a lab elled image (for V2) on the partitions in parallel. A novel sub-tree merging alg orithm is used in V1 to merge the sub-trees into the final tree, and the doubly linked lists are also merged in the process. While V2 merges the sub-trees using an existing merging algorithm. Finally, MSERs are recognized, the pixels in the m are extracted through two novel pixel extraction methods taking advantage of t he fact that a lot of pixels in parent and child MSERs are duplicated. Both V1 a nd V2 outperform three open source MSER algorithms (28 and 26 times faster than OpenCV MSER), and reduce the memory of the pixels in MSERs by 78%.
****
Seeing Around Street Corners: Non-Line-of-Sight Detection and Tracking In-the-Wi ld Using Doppler Radar
Nicolas Scheiner, Florian Kraus, Fangyin Wei, Buu Phan, Fahim Mannan, Nils Appenrodt, Werner Ritter, Jurgen Dickmann, Klaus Dietmayer, Bernhard Sick, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2020, pp. 2068-2077
Conventional sensor systems record information about directly visible objects, w hereas occluded scene components are considered lost in the measurement process. Non-line-of-sight (NLOS) methods try to recover such hidden objects from their indirect reflections - faint signal components, traditionally treated as measure ment noise. Existing NLOS approaches struggle to record these low-signal compone nts outside the lab, and do not scale to large-scale outdoor scenes and high-spe ed motion, typical in automotive scenarios. In particular, optical NLOS capture is fundamentally limited by the quartic intensity falloff of diffuse indirect re flections. In this work, we depart from visible-wavelength approaches and demons trate detection, classification, and tracking of hidden objects in large-scale d ynamic environments using Doppler radars that can be manufactured at low-cost in series production. To untangle noisy indirect and direct reflections, we learn from temporal sequences of Doppler velocity and position measurements, which we fuse in a joint NLOS detection and tracking network over time. We validate the a pproach on in-the-wild automotive scenes, including sequences of parked cars or house facades as relay surfaces, and demonstrate low-cost, real-time NLOS in dyn amic automotive environments.
****
Weakly Supervised Visual Semantic Parsing
Alireza Zareian, Svebor Karaman, Shih-Fu Chang; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3736-3745
Scene Graph Generation (SGG) aims to extract entities, predicates and their sema ntic structure from images, enabling deep understanding of visual content, with many applications such as visual reasoning and image retrieval. Nevertheless, ex isting SGG methods require millions of manually annotated bounding boxes for tra ining, and are computationally inefficient, as they exhaustively process all pai rs of object proposals to detect predicates. In this paper, we address those two limitations by first proposing a generalized formulation of SGG, namely Visual Semantic Parsing, which disentangles entity and predicate recognition, and enabl es sub-quadratic performance. Then we propose the Visual Semantic Parsing Networ k, VSPNet, based on a dynamic, attention-based, bipartite message passing framew ork that jointly infers graph nodes and edges through an iterative process. Addi

tionally, we propose the first graph-based weakly supervised learning framework, based on a novel graph alignment algorithm, which enables training without bounding box annotations. Through extensive experiments, we show that VSPNet outperforms weakly supervised baselines significantly and approaches fully supervised performance, while being several times faster. We publicly release the source code of our method.
****

Bringing Old Photos Back to Life
Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2747-2757
We propose to restore old photos that suffer from severe degradation through a deep learning approach. Unlike conventional restoration tasks that can be solved through supervised learning, the degradation in real photos is complex and the domain gap between synthetic images and real old photos makes the network fail to generalize. Therefore, we propose a novel triplet domain translation network by leveraging real photos along with massive synthetic image pairs. Specifically, we train two variational autoencoders (VAEs) to respectively transform old photos and clean photos into two latent spaces. And the translation between these two latent spaces is learned with synthetic paired data. This translation generalizes well to real photos because the domain gap is closed in the compact latent space. Besides, to address multiple degradations mixed in one old photo, we design a global branch with a partial nonlocal block targeting to the structured defects, such as scratches and dust spots, and a local branch targeting to the unstructured defects, such as noises and blurriness. Two branches are fused in the latent space, leading to improved capability to restore old photos from multiple defects. The proposed method outperforms state-of-the-art methods in terms of visual quality for old photos restoration.
****

Enhanced Blind Face Restoration With Multi-Exemplar Images and Adaptive Spatial Feature Fusion
Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2706-2715
In many real-world face restoration applications, e.g., smartphone photo albums and old films, multiple high-quality (HQ) images of the same person usually are available for a given degraded low-quality (LQ) observation. However, most existing guided face restoration methods are based on single HQ exemplar image, and are limited in properly exploiting guidance for improving the generalization ability to unknown degradation process. To address these issues, this paper suggests to enhance blind face restoration performance by utilizing multi-exemplar images and adaptive fusion of features from guidance and degraded images. First, given a degraded observation, we select the optimal guidance based on the weighted affine distance on landmark sets, where the landmark weights are learned to make the guidance image optimized to HQ image reconstruction. Second, moving least-square and adaptive instance normalization are leveraged for spatial alignment and illumination translation of guidance image in the feature space. Finally, for better feature fusion, multiple adaptive spatial feature fusion (ASFF) layers are introduced to incorporate guidance features in an adaptive and progressive manner, resulting in our ASFFNet. Experiments show that our ASFFNet performs favorably in terms of quantitative and qualitative evaluation, and is effective in generating photo-realistic results on real-world LQ images. The source code and models are available at https://github.com/csxmli2016/ASFFNet.
****

Geometry-Aware Satellite-to-Ground Image Synthesis for Urban Areas
Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, Rongjun Qin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 859-867
We present a novel method for generating panoramic street-view images which are geometrically consistent with a given satellite image. Different from existing a

pproaches that completely rely on a deep learning architecture to generalize cro ss-view image distributions, our approach explicitly loops in the geometric conf iguration of the ground objects based on the satellite views, such that the prod uced ground view synthesis preserves the geometric shape and the semantics of th e scene. In particular, we propose a neural network with a geo-transformation la yer that turns predicted ground-height values from the satellite view to a groun d view while retaining the physical satellite-to-ground relation. Our results sh ow that the synthesized image retains well-articulated and authentic geometric s hapes, as well as texture richness of the street-view in various scenarios. Both qualitative and quantitative results demonstrate that our method compares favor ably to other state-of-the-art approaches that lack geometric consistency.
****

End-to-End Learning Local Multi-View Descriptors for 3D Point Clouds
Lei Li, Siyu Zhu, Hongbo Fu, Ping Tan, Chiew-Lan Tai; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 19 19-1928
In this work, we propose an end-to-end framework to learn local multi-view descr iptors for 3D point clouds. To adopt a similar multi-view representation, existi ng studies use hand-crafted viewpoints for rendering in a preprocessing stage, w hich is detached from the subsequent descriptor learning stage. In our framework , we integrate the multi-view rendering into neural networks by using a differen tiable renderer, which allows the viewpoints to be optimizable parameters for ca pturing more informative local context of interest points. To obtain discriminat ive descriptors, we also design a soft-view pooling module to attentively fuse c onvolutional features across views. Extensive experiments on existing 3D registr ation benchmarks show that our method outperforms existing local descriptors bot h quantitatively and qualitatively.
****

Multi-Scale Boosted Dehazing Network With Dense Feature Fusion
Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, Ming-Hsu an Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2020, pp. 2157-2167
In this paper, we propose a Multi-Scale Boosted Dehazing Network with Dense Feat ure Fusion based on the U-Net architecture. The proposed method is designed base d on two principles, boosting and error feedback, and we show that they are suit able for the dehazing problem. By incorporating the Strengthen-Operate-Subtract boosting strategy in the decoder of the proposed model, we develop a simple yet effective boosted decoder to progressively restore the haze-free image. To addre ss the issue of preserving spatial information in the U-Net architecture, we des ign a dense feature fusion module using the back-projection feedback scheme. We show that the dense feature fusion module can simultaneously remedy the missing spatial information from high-resolution features and exploit the non-adjacent f eatures. Extensive evaluations demonstrate that the proposed model performs favo rably against the state-of-the-art approaches on the benchmark datasets as well as real-world hazy images.
****

A Multi-Hypothesis Approach to Color Constancy
Daniel Hernandez-Juarez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Greg ory Slabaugh, Steven McDonagh; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2020, pp. 2270-2280
Contemporary approaches frame the color constancy problem as learning camera spe cific illuminant mappings. While high accuracy can be achieved on camera specifi c data, these models depend on camera spectral sensitivity and typically exhibit poor generalisation to new devices. Additionally, regression methods produce po int estimates that do not explicitly account for potential ambiguities among pla usible illuminant solutions, due to the ill-posed nature of the problem. We prop ose a Bayesian framework that naturally handles color constancy ambiguity via a multi-hypothesis strategy. Firstly, we select a set of candidate scene illuminan ts in a data-driven fashion and apply them to a target image to generate a set o f corrected images. Secondly, we estimate, for each corrected image, the likelih

ood of the light source being achromatic using a camera-agnostic CNN. Finally, o ur method explicitly learns a final illumination estimate from the generated pos terior probability distribution. Our likelihood estimator learns to answer a cam era-agnostic question and thus enables effective multi-camera training by disent angling illuminant estimation from the supervised learning task. We extensively evaluate our proposed approach and additionally set a benchmark for novel sensor generalisation without re-training. Our method provides state-of-the-art accura cy on multiple public datasets (up to 11% median angular error improvement) whil e maintaining real-time execution.
****

MSeg: A Composite Dataset for Multi-Domain Semantic Segmentation
John Lambert, Zhuang Liu, Ozan Sener, James Hays, Vladlen Koltun; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2879-2888
We present MSeg, a composite dataset that unifies se- mantic segmentation datase ts from different domains. A naive merge of the constituent datasets yields poor performance due to inconsistent taxonomies and annotation practices. We reconci le the taxonomies and bring the pixel-level annotations into alignment by relabe ling more than 220,000 object masks in more than 80,000 images. The resulting co mposite dataset enables training a single semantic segmentation model that funct ions effectively across domains and generalizes to datasets that were not seen d uring training. We adopt zero-shot cross-dataset transfer as a benchmark to syst ematically evaluate a model's robustness and show that MSeg training yields subs tantially more robust models in comparison to training on individual datasets or naive mixing of datasets without the presented contributions. A model trained o n MSeg ranks first on the WildDash leaderboard for robust semantic segmentation, with no exposure to WildDash data during training.
****

Learning Event-Based Motion Deblurring
Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, Yebin Liu; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2020, pp. 3320-3329
Recovering sharp video sequence from a motion-blurred image is highly ill-posed due to the significant loss of motion information in the blurring process. For e vent-based cameras, however, fast motion can be captured as events at high frame rate, raising new opportunities to exploring effective solutions. In this paper , we start from a sequential formulation of event-based motion deblurring, then show how its optimization can be unfolded with a novel end-toend deep architectu re. The proposed architecture is a convolutional recurrent neural network that i ntegrates visual and temporal knowledge of both global and local scales in princ ipled manner. To further improve the reconstruction, we propose a differentiable directional event filtering module to effectively extract rich boundary prior f rom the evolution of events. We conduct extensive experiments on the synthetic G oPro dataset and a large newly introduced dataset captured by a DAVIS240C camera . The proposed approach achieves state-of-the-art reconstruction quality, and ge neralizes better to handling real-world motion blur.
****

DMCP: Differentiable Markov Channel Pruning for Neural Networks
Shaopeng Guo, Yujie Wang, Quanquan Li, Junjie Yan; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1539-1 547
Recent works imply that the channel pruning can be regarded as searching optimal sub-structure from unpruned networks. However, existing works based on this obs ervation require training and evaluating a large number of structures, which lim its their application. In this paper, we propose a novel differentiable method f or channel pruning, named Differentiable Markov Channel Pruning (DMCP), to effic iently search the optimal sub-structure. Our method is differentiable and can be directly optimized by gradient descent with respect to standard task loss and b udget regularization (e.g. FLOPs constraint). In DMCP, we model the channel prun ing as a Markov process, in which each state represents for retaining the corres

ponding channel during pruning, and transitions between states denote the pruning process. In the end, our method is able to implicitly select the proper number of channels in each layer by the Markov process with optimized transitions. To validate the effectiveness of our method, we perform extensive experiments on Imagenet with ResNet and MobilenetV2. Results show our method can achieve consistent improvement than state-of-the-art pruning methods in various FLOPs settings.
****

From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement
Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, Jiaying Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3063-3072
Under-exposure introduces a series of visual degradation, i.e. decreased visibility, intensive noise, and biased color, etc. To address these problems, we propose a novel semi-supervised learning approach for low-light image enhancement. A deep recursive band network (DRBN) is proposed to recover a linear band representation of an enhanced normal-light image with paired low/normal-light images, and then obtain an improved one by recomposing the given bands via another learnable linear transformation based on a perceptual quality-driven adversarial learning with unpaired data. The architecture is powerful and flexible to have the merit of training with both paired and unpaired data. On one hand, the proposed network is well designed to extract a series of coarse-to-fine band representations, whose estimations are mutually beneficial in a recursive process. On the other hand, the extracted band representation of the enhanced image in the first stage of DRBN (recursive band learning) bridges the gap between the restoration knowledge of paired data and the perceptual quality preference to real high-quality images. Its second stage (band recomposition) learns to recompose the band representation towards fitting perceptual properties of high-quality images via adversarial learning. With the help of this two-stage design, our approach generates enhanced results with well-reconstructed details and visually promising contrast and color distributions. Qualitative and quantitative evaluations demonstrate the superiority of our DRBN.
****

Learning Integral Objects With Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation
Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4283-4292
Image-level weakly-supervised semantic segmentation (WSSS) aims at learning semantic segmentation by adopting only image class labels. Existing approaches generally rely on class activation maps (CAM) to generate pseudo-masks and then train segmentation models. The main difficulty is that the CAM estimate only covers partial foreground objects. In this paper, we argue that the critical factor preventing to obtain the full object mask is the classification boundary mismatch problem in applying the CAM to WSSS. Because the CAM is optimized by the classification task, it focuses on the discrimination across different image-level classes. However, the WSSS requires to distinguish pixels sharing the same image-level class to separate them into the foreground and the background. To alleviate this contradiction, we propose an efficient end-to-end Intra-Class Discriminator (ICD) framework, which learns intra-class boundaries to help separate the foreground and the background within each image-level class. Without bells and whistles, our approach achieves the state-of-the-art performance of image label based WSSS, with mIoU 68.0% on the VOC 2012 semantic segmentation benchmark, demonstrating the effectiveness of the proposed approach.
****

Enhancing Cross-Task Black-Box Transferability of Adversarial Examples With Dispersion Reduction
Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, Senem Velipasalar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 940-949

Neural networks are known to be vulnerable to carefully crafted adversarial exam
ples, and these malicious samples often transfer, i.e., they remain adversarial
even against other models. Although significant effort has been devoted to the t
ransferability across models, surprisingly little attention has been paid to cro
ss-task transferability, which represents the real-world cybercriminal's situati
on, where an ensemble of different defense/detection mechanisms need to be evade
d all at once. We investigate the transferability of adversarial examples across
 a wide range of real-world computer vision tasks, including image classificatio
n, object detection, semantic segmentation, explicit content detection, and text
 detection. Our proposed attack minimizes the "dispersion" of the internal featu
re map, overcoming the limitations of existing attacks, that require task-specif
ic loss functions and/or probing a target model. We conduct evaluation on open-s
ource detection and segmentation models, as well as four different computer visi
on tasks provided by Google Cloud Vision (GCV) APIs. We demonstrate that our app
roach outperforms existing attacks by degrading performance of multiple CV tasks
 by a large margin with only modest perturbations.
****

Mesh-Guided Multi-View Stereo With Pyramid Architecture
Yuesong Wang, Tao Guan, Zhuo Chen, Yawei Luo, Keyang Luo, Lili Ju; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2020, pp. 2039-2048
Multi-view stereo (MVS) aims to reconstruct 3D geometry of the target scene by u
sing only information from 2D images. Although much progress has been made, it s
till suffers from textureless regions. To overcome this difficulty, we propose a
 mesh-guided MVS method with pyramid architecture, which makes use of the surfac
e mesh obtained from coarse-scale images to guide the reconstruction process. Sp
ecifically, a PatchMatch-based MVS algorithm is first used to generate depth map
s for coarse-scale images and the corresponding surface mesh is obtained by a su
rface reconstruction algorithm. Next we project the mesh onto each of depth maps
 to replace unreliable depth values and the corrected depth maps are fed to fine
-scale reconstruction for initialization. To alleviate the influence of possible
 erroneous faces on the mesh, we further design and train a convolutional neural
 network to remove incorrect depths. In addition, it is often hard for the corre
ct depth values for low-textured regions to survive at the fine-scale, thus we a
lso develop an efficient method to seek out these regions and further enforce th
e geometric consistency in these regions. Experimental results on the ETH3D high
-resolution dataset demonstrate that our method achieves state-of-the-art perfor
mance, especially in completeness.
****

BSP-Net: Generating Compact Meshes via Binary Space Partitioning
Zhiqin Chen, Andrea Tagliasacchi, Hao Zhang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 45-54
Polygonal meshes are ubiquitous in the digital 3D domain, yet they have only pla
yed a minor role in the deep learning revolution. Leading methods for learning g
enerative models of shapes rely on implicit functions, and generate meshes only
 after expensive iso-surfacing routines. To overcome these challenges, we are ins
pired by a classical spatial data structure from computer graphics, Binary Space
 Partitioning (BSP), to facilitate 3D learning. The core ingredient of BSP is an
 operation for recursive subdivision of space to obtain convex sets. By exploiti
ng this property, we devise BSP-Net, a network that learns to represent a 3D sha
pe via convex decomposition. Importantly, BSP-Net is unsupervised since no conve
x shape decompositions are needed for training. The network is trained to recons
truct a shape using a set of convexes obtained from a BSP-tree built on a set of
 planes. The convexes inferred by BSP-Net can be easily extracted to form a poly
gon mesh, without any need for iso-surfacing. The generated meshes are compact (
i.e., low-poly) and well suited to represent sharp geometry; they are guaranteed
 to be watertight and can be easily parameterized. We also show that the reconst
ruction quality by BSP-Net is competitive with state-of-the-art methods while us
ing much fewer primitives. Code is available at https://github.com/czq142857/BSP
-NET-original.

****
Which Is Plagiarism: Fashion Image Retrieval Based on Regional Representation for Design Protection
Yining Lang, Yuan He, Fan Yang, Jianfeng Dong, Hui Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2595-2604
With the rapid growth of e-commerce and the popularity of online shopping, fashion retrieval has received considerable attention in the computer vision community. Different from the existing works that mainly focus on identical or similar fashion item retrieval, in this paper, we aim to study the plagiarized clothes retrieval which is somewhat ignored in the academic community while itself has great application value. One of the key challenges is that plagiarized clothes are usually modified in a certain region on the original design to escape the supervision by traditional retrieval methods. To relieve it, we propose a novel network named Plagiarized-Search-Net (PS-Net) based on regional representation, where we utilize the landmarks to guide the learning of regional representations and compare fashion items region by region. Besides, we propose a new dataset named Plagiarized Fashion for plagiarized clothes retrieval, which provides a meaningful complement to the existing fashion retrieval field. Experiments on Plagiarized Fashion dataset verify that our approach is superior to other instance-level counterparts for plagiarized clothes retrieval, showing a promising result for original design protection. Moreover, our PS-Net can also be adapted to traditional fashion retrieval and landmark estimation tasks and achieves the state-of-the-art performance on the DeepFashion and DeepFashion2 datasets.
****
Relation-Aware Global Attention for Person Re-Identification
Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3186-3195
For person re-identification (re-id), attention mechanisms have become attractive as they aim at strengthening discriminative features and suppressing irrelevant ones, which matches well the key of re-id, i.e., discriminative feature learning. Previous approaches typically learn attention using local convolutions, ignoring the mining of knowledge from global structure patterns. Intuitively, the affinities among spatial positions/nodes in the feature map provide clustering-like information and are helpful for inferring semantics and thus attention, especially for person images where the feasible human poses are constrained. In this work, we propose an effective Relation-Aware Global Attention (RGA) module which captures the global structural information for better attention learning. Specifically, for each feature position, in order to compactly grasp the structural information of global scope and local appearance information, we propose to stack the relations, i.e., its pairwise correlations/affinities with all the feature positions (e.g., in raster scan order), and the feature itself together to learn the attention with a shallow convolutional model. Extensive ablation studies demonstrate that our RGA can significantly enhance the feature representation power and help achieve the state-of-the-art performance on several popular benchmarks. The source code is available at https://github.com/microsoft/Relation-Aware-Global-Attention-Networks.
****
Adversarial Camouflage: Hiding Physical-World Attacks With Natural Styles
Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. K. Qin, Yun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1000-1008
Deep neural networks (DNNs) are known to be vulnerable to adversarial examples. Existing works have mostly focused on either digital adversarial examples created via small and imperceptible perturbations, or physical-world adversarial examples created with large and less realistic distortions that are easily identified by human observers. In this paper, we propose a novel approach, called Adversarial Camouflage (AdvCam), to craft and camouflage physical-world adversarial examples into natural styles that appear legitimate to human observers. Specifically

, AdvCam transfers large adversarial perturbations into customized styles, which are then "hidden" on-target object or off-target background. Experimental evaluation shows that, in both digital and physical-world scenarios, adversarial examples crafted by AdvCam are well camouflaged and highly stealthy, while remaining effective in fooling state-of-the-art DNN image classifiers. Hence, AdvCam is a flexible approach that can help craft stealthy attacks to evaluate the robustness of DNNs.
****
Evolving Losses for Unsupervised Video Representation Learning
AJ Piergiovanni, Anelia Angelova, Michael S. Ryoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 133-142
We present a new method to learn video representations from large-scale unlabeled video data. Ideally, this representation will be generic and transferable, directly usable for new tasks such as action recognition and zero or few-shot learning. We formulate unsupervised representation learning as a multi-modal, multi-task learning problem, where the representations are shared across different modalities via distillation. Further, we introduce the concept of loss function evolution by using an evolutionary search algorithm to automatically find optimal combination of loss functions capturing many (self-supervised) tasks and modalities. Thirdly, we propose an unsupervised representation evaluation metric using distribution matching to a large unlabeled dataset as a prior constraint, based on Zipf's law. This unsupervised constraint, which is not guided by any labeling, produces similar results to weakly-supervised, task-specific ones. The proposed unsupervised representation learning results in a single RGB network and outperforms previous methods. Notably, it is also more effective than several label-based methods (e.g., ImageNet), with the exception of large, fully labeled video datasets.
****
3DV: 3D Dynamic Voxel for Action Recognition in Depth Video
Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, Junsong Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 511-520
For depth-based 3D action recognition, one essential issue is to represent 3D motion pattern effectively and efficiently. To this end, 3D dynamic voxel (3DV) is proposed as a novel 3D motion representation manner. With 3D space voxelization, the key idea of 3DV is to encode the 3D motion information within depth video into a regular voxel set (i.e., 3DV) compactly, via temporal rank pooling. Each available 3DV voxel intrinsically involves 3D spatial and motion feature for 3D action description. 3DV is then abstracted as a point set and input into PointNet++ for 3D action recognition, in the end-to-end learning way. The intuition for transferring 3DV into the point set form is that, PointNet++ is lightweight and effective for deep feature learning towards point set. Since 3DV may loose appearance clue, a multi-stream 3D action recognition manner is also proposed to learn motion and appearance feature jointly. To extract richer temporal order information of actions, we also split the depth video into temporal segments and encode this procedure in 3DV integrally. The extensive experiments on the well-established benchmark datasets (e.g., NTU RGB+D 120 and NTU RGB+D 60) demonstrate the superiority of our proposition. Impressively, we acquire the accuracy of 82.4% and 93.5% on NTU RGB+D 120 with the cross-subject and cross-setup test setting respectively. 3DV's code is available at https://github.com/3huo/3DV-Action.
****
Benchmarking Adversarial Robustness on Image Classification
Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 321-331
Deep neural networks are vulnerable to adversarial examples, which becomes one of the most important research problems in the development of deep learning. While a lot of efforts have been made in recent years, it is of great significance to perform correct and complete evaluations of the adversarial attack and defense algorithms. In this paper, we establish a comprehensive, rigorous, and coherent

benchmark to evaluate adversarial robustness on image classification tasks. After briefly reviewing plenty of representative attack and defense methods, we perform large-scale experiments with two robustness curves as the fair-minded evaluation criteria to fully understand the performance of these methods. Based on the evaluation results, we draw several important findings that can provide insights for future research, including: 1) The relative robustness between models can change across different attack configurations, thus it is encouraged to adopt the robustness curves to evaluate adversarial robustness; 2) As one of the most effective defense techniques, adversarial training can generalize across different threat models; 3) Randomization-based defenses are more robust to query-based black-box attacks.

****

What It Thinks Is Important Is Important: Robustness Transfers Through Input Gradients

Alvin Chan, Yi Tay, Yew-Soon Ong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 332-341

Adversarial perturbations are imperceptible changes to input pixels that can change the prediction of deep learning models. Learned weights of models robust to such perturbations are previously found to be transferable across different tasks but this applies only if the model architecture for the source and target tasks is the same. Input gradients characterize how small changes at each input pixel affect the model output. Using only natural images, we show here that training a student model's input gradients to match those of a robust teacher model can gain robustness close to a strong baseline that is robustly trained from scratch. Through experiments in MNIST, CIFAR-10, CIFAR-100 and Tiny-ImageNet, we show that our proposed method, input gradient adversarial matching, can transfer robustness across different tasks and even across different model architectures. This demonstrates that directly targeting the semantics of input gradients is a feasible way towards adversarial robustness.

****

Gum-Net: Unsupervised Geometric Matching for Fast and Accurate 3D Subtomogram Image Alignment and Averaging

Xiangrui Zeng, Min Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4073-4084

We propose a Geometric unsupervised matching Net-work (Gum-Net) for finding the geometric correspondence between two images with application to 3D subtomogram alignment and averaging. Subtomogram alignment is the most important task in cryo-electron tomography (cryo-ET), a revolutionary 3D imaging technique for visualizing the molecular organization of unperturbed cellular landscapes in single cells. However, subtomogram alignment and averaging are very challenging due to severe imaging limits such as noise and missing wedge effects. We introduce an end-to-end trainable architecture with three novel modules specifically designed for preserving feature spatial information and propagating feature matching information. The training is performed in a fully unsupervised fashion to optimize a matching metric. No ground truth transformation information nor category-level or instance-level matching supervision information is needed. After systematic assessments on six real and nine simulated datasets, we demonstrate that Gum-Net reduced the alignment error by 40 to 50% and improved the averaging resolution by 10%. Gum-Net also achieved 70 to 110 times speedup in practice with GPU acceleration compared to state-of-the-art subtomogram alignment methods. Our work is the first 3D unsupervised geometric matching method for images of strong transformation variation and high noise level. The training code, trained model, and datasets are available in our open-source software AITom.

****

Deep Parametric Shape Predictions Using Distance Fields

Dmitriy Smirnov, Matthew Fisher, Vladimir G. Kim, Richard Zhang, Justin Solomon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 561-570

Many tasks in graphics and vision demand machinery for converting shapes into consistent representations with sparse sets of parameters; these representations f

acilitate rendering, editing, and storage. When the source data is noisy or ambi guous, however, artists and engineers often manually construct such representati ons, a tedious and potentially time-consuming process. While advances in deep le arning have been successfully applied to noisy geometric data, the task of gener ating parametric shapes has so far been difficult for these methods. Hence, we p ropose a new framework for predicting parametric shape primitives using deep lea rning. We use distance fields to transition between shape parameters like contro l points and input data on a pixel grid. We demonstrate efficacy on 2D and 3D ta sks, including font vectorization and surface abstraction.
****

Correspondence-Free Material Reconstruction using Sparse Surface Constraints
Sebastian Weiss, Robert Maier, Daniel Cremers, Rudiger Westermann, Nils Thue rey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2020, pp. 4686-4695
We present a method to infer physical material parameters, and even external bou ndaries, from the scanned motion of a homogeneous deformable object via the solu tion of an inverse problem. Parameters are estimated from real-world data source s such as sparse observations from a Kinect sensor without correspondences. We i ntroduce a novel Lagrangian-Eulerian optimization formulation, including a cost function that penalizes differences to observations during an optimization run. This formulation matches correspondence-free, sparse observations from a single- view depth image with a finite element simulation of deformable bodies. In a num ber of tests using synthetic datasets and real-world measurements, we analyse th e robustness of our approach and the convergence behavior of the numerical optim ization scheme.
****

PointPainting: Sequential Fusion for 3D Object Detection
Sourabh Vora, Alex H. Lang, Bassam Helou, Oscar Beijbom; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4604-4612
Camera and lidar are important sensor modalities for robotics in general and sel f-driving cars in particular. The sensors provide complementary information offe ring an opportunity for tight sensor-fusion. Surprisingly, lidar-only methods ou tperform fusion methods on the main benchmark datasets, suggesting a gap in the literature. In this work, we propose PointPainting: a sequential fusion method t o fill this gap. PointPainting works by projecting lidar points into the output of an image-only semantic segmentation network and appending the class scores to each point. The appended (painted) point cloud can then be fed to any lidar-onl y method. Experiments show large improvements on three different state-of-the ar t methods, Point-RCNN, VoxelNet and PointPillars on the KITTI and nuScenes datas ets. The painted version of PointRCNN represents a new state of the art on the K ITTI leaderboard for the bird's-eye view detection task. In ablation, we study h ow the effects of Painting depends on the quality and format of the semantic seg mentation output, and demonstrate how latency can be minimized through pipelinin g.
****

Adaptive Subspaces for Few-Shot Learning
Christian Simon, Piotr Koniusz, Richard Nock, Mehrtash Harandi; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 0, pp. 4136-4145
Object recognition requires a generalization capability to avoid overfitting, es pecially when the samples are extremely few. Generalization from limited samples , usually studied under the umbrella of meta-learning, equips learning technique s with the ability to adapt quickly in dynamical environments and proves to be a n essential aspect of life long learning. In this paper, we provide a framework for few-shot learning by introducing dynamic classifiers that are constructed fr om few samples. A subspace method is exploited as the central block of a dynamic classifier. We will empirically show that such modelling leads to robustness ag ainst perturbations (e.g., outliers) and yields competitive results on the task of supervised and semi-supervised few-shot classification. We also develop a dis

criminative form which can boost the accuracy even further. Our code is availabl
e at https://github.com/chrysts/dsn_fewshot
****

In Perfect Shape: Certifiably Optimal 3D Shape Reconstruction From 2D Landmarks
Heng Yang, Luca Carlone; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2020, pp. 621-630
We study the problem of 3D shape reconstruction from 2D landmarks extracted in a
 single image. We adopt the 3D deformable shape model and formulate the reconstr
uction as a joint optimization of the camera pose and the linear shape parameter
s. Our first contribution is to apply Lasserre's hierarchy of convex Sums-of-Squ
ares (SOS) relaxations to solve the shape reconstruction problem and show that t
he SOS relaxation of minimum order 2 empirically solves the original non-convex
problem exactly. Our second contribution is to exploit the structure of the poly
nomial in the objective function and find a reduced set of basis monomials for t
he SOS relaxation that significantly decreases the size of the resulting semidef
inite program (SDP) without compromising its accuracy. These two contributions,
to the best of our knowledge, lead to the first certifiably optimal solver for 3
D shape reconstruction, that we name Shape*. Our third contribution is to add an
 outlier rejection layer to Shape[?] using a truncated least squares (TLS) robus
t cost function and leveraging graduated non-convexity to solve TLS without init
ialization. The result is a robust reconstruction algorithm, named Shape#, that
tolerates a large amount of outlier measurements. We evaluate the performance of
 Shape[?] and Shape# in both simulated and real experiments, showing that Shape[
?] outperforms local optimization and previous convex relaxation techniques, whi
le Shape# achieves state-of-the-art performance and is robust against 70% outlie
rs in the FG3DCar dataset.
****

Dynamic Graph Message Passing Networks
Li Zhang, Dan Xu, Anurag Arnab, Philip H.S. Torr; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3726-37
35
Modelling long-range dependencies is critical for scene understanding tasks in c
omputer vision. Although CNNs have excelled in many vision tasks, they are still
 limited in capturing long-range structured relationships as they typically cons
ist of layers of local kernels. A fully-connected graph is beneficial for such m
odelling, however, its computational overhead is prohibitive. We propose a dynam
ic graph message passing network, that significantly reduces the computational c
omplexity compared to related works modelling a fully-connected graph. This is a
chieved by adaptively sampling nodes in the graph, conditioned on the input, for
 message passing. Based on the sampled nodes, we dynamically predict node-depend
ent filter weights and the affinity matrix for propagating information between t
hem. Using this model, we show significant improvements with respect to strong,
state-of-the-art baselines on three different tasks and backbone architectures.
Our approach also outperforms fully-connected graphs while using substantially f
ewer floating-point operations and parameters.
****

Evaluating Weakly Supervised Object Localization Methods Right
Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, Hyunj
ung Shim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2020, pp. 3133-3142
Weakly-supervised object localization (WSOL) has gained popularity over the last
 years for its promise to train localization models with only image-level labels
. Since the seminal WSOL work of class activation mapping (CAM), the field has f
ocused on how to expand the attention regions to cover objects more broadly and
localize them better. However, these strategies rely on full localization superv
ision to validate hyperparameters and for model selection, which is in principle
 prohibited under the WSOL setup. In this paper, we argue that WSOL task is ill-
posed with only image-level labels, and propose a new evaluation protocol where
full supervision is limited to only a small held-out set not overlapping with th
e test set. We observe that, under our protocol, the five most recent WSOL metho

ds have not made a major improvement over the CAM baseline. Moreover, we report that existing WSOL methods have not reached the few-shot learning baseline, where the full-supervision at validation time is used for model training instead. Based on our findings, we discuss some future directions for WSOL.
****
ClusterVO: Clustering Moving Instances and Estimating Visual Odometry for Self and Surroundings
Jiahui Huang, Sheng Yang, Tai-Jiang Mu, Shi-Min Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2168-2177
We present ClusterVO, a stereo Visual Odometry which simultaneously clusters and estimates the motion of both ego and surrounding rigid clusters/objects. Unlike previous solutions relying on batch input or imposing priors on scene structure or dynamic object models, ClusterVO is online, general and thus can be used in various scenarios including indoor scene understanding and autonomous driving. At the core of our system lies a multi-level probabilistic association mechanism and a heterogeneous Conditional Random Field (CRF) clustering approach combining semantic, spatial and motion information to jointly infer cluster segmentations online for every frame. The poses of camera and dynamic objects are instantly solved through a sliding-window optimization. Our system is evaluated on Oxford Multimotion and KITTI dataset both quantitatively and qualitatively, reaching comparable results to state-of-the-art solutions on both odometry and dynamic trajectory recovery.
****
DaST: Data-Free Substitute Training for Adversarial Attacks
Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, Ce Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 234-243
Machine learning models are vulnerable to adversarial examples. For the black-box setting, current substitute attacks need pre-trained models to generate adversarial examples. However, pre-trained models are hard to obtain in real-world tasks. In this paper, we propose a data-free substitute training method (DaST) to obtain substitute models for adversarial black-box attacks without the requirement of any real data. To achieve this, DaST utilizes specially designed generative adversarial networks (GANs) to train the substitute models. In particular, we design a multi-branch architecture and label-control loss for the generative model to deal with the uneven distribution of synthetic samples. The substitute model is then trained by the synthetic samples generated by the generative model, which are labeled by the attacked model subsequently. The experiments demonstrate the substitute models produced by DaST can achieve competitive performance compared with the baseline models which are trained by the same train set with attacked models. Additionally, to evaluate the practicability of the proposed method on the real-world task, we attack an online machine learning model on the Microsoft Azure platform. The remote model misclassifies 98.35% of the adversarial examples crafted by our method. To the best of our knowledge, we are the first to train a substitute model for adversarial attacks without any real data.
****
Cross-Domain Semantic Segmentation via Domain-Invariant Interactive Relation Transfer
Fengmao Lv, Tao Liang, Xiang Chen, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4334-4343
Exploiting photo-realistic synthetic data to train semantic segmentation models has received increasing attention over the past years. However, the domain mismatch between synthetic and real images will cause a significant performance drop when the model trained with synthetic images is directly applied to real-world scenarios. In this paper, we propose a new domain adaptation approach, called Pivot Interaction Transfer (PIT). Our method mainly focuses on constructing pivot information that is common knowledge shared across domains as a bridge to promote the adaptation of semantic segmentation model from synthetic domains to real-wo

rld domains. Specifically, we first infer the image-level category information a bout the target images, which is then utilized to facilitate pixel-level transfe r for semantic segmentation, with the assumption that the interactive relation b etween the image-level category information and the pixel-level semantic informa tion is invariant across domains. To this end, we propose a novel multi-level re gion expansion mechanism that aligns both the image-level and pixel-level inform ation. Comprehensive experiments on the adaptation from both GTAV and SYNTHIA to Cityscapes clearly demonstrate the superiority of our method.
****

Minimal Solutions for Relative Pose With a Single Affine Correspondence
Banglei Guan, Ji Zhao, Zhang Li, Fang Sun, Friedrich Fraundorfer; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1929-1938
In this paper we present four cases of minimal solutions for two-view relative p ose estimation by exploiting the affine transformation between feature points an d we demonstrate efficient solvers for these cases. It is shown, that under the planar motion assumption or with knowledge of a vertical direction, a single aff ine correspondence is sufficient to recover the relative camera pose. The four c ases considered are two-view planar relative motion for calibrated cameras as a closed-form and a least-squares solution, a closed-form solution for unknown foc al length and the case of a known vertical direction. These algorithms can be us ed efficiently for outlier detection within a RANSAC loop and for initial motion estimation. All the methods are evaluated on both synthetic data and real-world datasets from the KITTI benchmark. The experimental results demonstrate that ou r methods outperform comparable state-of-the-art methods in accuracy with the be nefit of a reduced number of needed RANSAC iterations.
****

Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analy sis
Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G. Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, David Wilson, Adrian Tudor, Naira Hovakimyan, Thomas S. Huang, Honghu i Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2020, pp. 2828-2838
The success of deep learning in visual recognition tasks has driven advancements in multiple fields of research. Particularly, increasing attention has been dra wn towards its application in agriculture. Nevertheless, while visual pattern re cognition on farmlands carries enormous economic values, little progress has bee n made to merge computer vision and crop sciences due to the lack of suitable ag ricultural image datasets. Meanwhile, problems in agriculture also pose new chal lenges in computer vision. For example, semantic segmentation of aerial farmland images requires inference over extremely large-size images with extreme annotat ion sparsity. These challenges are not present in most of the common object data sets, and we show that they are more challenging than many other aerial image da tasets. To encourage research in computer vision for agriculture, we present Agr iculture-Vision: a large-scale aerial farmland image dataset for semantic segmen tation of agricultural patterns. We collected 94,986 high-quality aerial images from 3,432 farmlands across the US, where each image consists of RGB and Near-in frared (NIR) channels with resolution as high as 10 cm per pixel. We annotate ni ne types of field anomaly patterns that are most important to farmers. As a pilo t study of aerial agricultural semantic segmentation, we perform comprehensive e xperiments using popular semantic segmentation models; we also propose an effect ive model designed for aerial agricultural pattern recognition. Our experiments demonstrate several challenges Agriculture-Vision poses to both the computer vis ion and agriculture communities. Future versions of this dataset will include ev en more aerial images, anomaly patterns and image channels.
****

ActiveMoCap: Optimized Viewpoint Selection for Active Human Motion Capture
Sena Kiciroglu, Helge Rhodin, Sudipta N. Sinha, Mathieu Salzmann, Pascal Fua ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit

ion (CVPR), 2020, pp. 103-112
The accuracy of monocular 3D human pose estimation depends on the viewpoint from which the image is captured. While freely moving cameras, such as on drones, provide control over this viewpoint, automatically positioning them at the location which will yield the highest accuracy remains an open problem. This is the problem that we address in this paper. Specifically, given a short video sequence, we introduce an algorithm that predicts which viewpoints should be chosen to capture future frames so as to maximize 3D human pose estimation accuracy. The key idea underlying our approach is a method to estimate the uncertainty of the 3D body pose estimates. We integrate several sources of uncertainty, originating from deep learning based regressors and temporal smoothness. Our motion planner yields improved 3D body pose estimates and outperforms or matches existing ones that are based on person following and orbiting.
****

ScrabbleGAN: Semi-Supervised Varying Length Handwritten Text Generation
Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, Roee Litman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4324-4333
Optical character recognition (OCR) systems performance have improved significantly in the deep learning era. This is especially true for handwritten text recognition (HTR), where each author has a unique style, unlike printed text, where the variation is smaller by design. That said, deep learning based HTR is limited, as in every other task, by the number of training examples. Gathering data is a challenging and costly task, and even more so, the labeling task that follows, of which we focus here. One possible approach to reduce the burden of data annotation is semi-supervised learning. Semi supervised methods use, in addition to labeled data, some unlabeled samples to improve performance, compared to fully supervised ones. Consequently, such methods may adapt to unseen images during test time. We present ScrabbleGAN, a semi-supervised approach to synthesize handwritten text images that are versatile both in style and lexicon. ScrabbleGAN relies on a novel generative model which can generate images of words with an arbitrary length. We show how to operate our approach in a semi-supervised manner, enjoying the aforementioned benefits such as performance boost over state of the art supervised HTR. Furthermore, our generator can manipulate the resulting text style. This allows us to change, for instance, whether the text is cursive, or how thin is the pen stroke.
****

Scalability in Perception for Autonomous Driving: Waymo Open Dataset
Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2446-2454
The research community has increasing interest in autonomous driving research, despite the resource intensity of obtaining representative real world data. Existing self-driving datasets are limited in the scale and variation of the environments they capture, even though generalization within and between operating regions is crucial to the over-all viability of the technology. In an effort to help align the research community's contributions with real-world self-driving problems, we introduce a new large scale, high quality, diverse dataset. Our new dataset consists of 1150 scenes that each span 20 seconds, consisting of well synchronized and calibrated high quality LiDAR and camera data captured across a range of urban and suburban geographies. It is 15x more diverse than the largest camera+LiDAR dataset available based on our proposed diversity metric. We exhaustively annotated this data with 2D (camera image) and 3D (LiDAR) bounding boxes, with consistent identifiers across frames. Finally, we provide strong baselines for 2D as well as 3D detection and tracking tasks. We further study the effects of dataset size and generalization across geographies on 3D detection methods. Find data, code and more up-to-date information at http://www.waymo.com/open.

****
Multi-Modal Domain Adaptation for Fine-Grained Action Recognition
Jonathan Munro, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 122-132
Fine-grained action recognition datasets exhibit environmental bias, where multiple video sequences are captured from a limited number of environments. Training a model in one environment and deploying in another results in a drop in performance due to an unavoidable domain shift. Unsupervised Domain Adaptation (UDA) approaches have frequently utilised adversarial training between the source and target domains. However, these approaches have not explored the multi-modal nature of video within each domain. In this work we exploit the correspondence of modalities as a self-supervised alignment approach for UDA in addition to adversarial alignment (Fig. 1). We test our approach on three kitchens from the large-scale EPIC-Kitchens dataset, using two modalities commonly employed for action recognition: RGB and Optical Flow. We show that multi-modal self-supervision alone improves the performance over source-only training by 2.4% on average. We then combine adversarial training with multi-modal self-supervision, showing that our approach outperforms other UDA methods by 3%.
****
A Sparse Resultant Based Method for Efficient Minimal Solvers
Snehal Bhayani, Zuzana Kukelova, Janne Heikkila; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1770-1779
Many computer vision applications require robust and efficient estimation of camera geometry. The robust estimation is usually based on solving camera geometry problems from a minimal number of input data measurements, i.e. solving minimal problems in a RANSAC framework. Minimal problems often result in complex systems of polynomial equations. Many state-of-the-art efficient polynomial solvers to these problems are based on Grobner basis and the action-matrix method that has been automatized and highly optimized in recent years. In this paper we study an alternative algebraic method for solving systems of polynomial equations, i.e., the sparse resultant-based method and propose a novel approach to convert the resultant constraint to an eigenvalue problem. This technique can significantly improve the efficiency and stability of existing resultant-based solvers. We applied our new resultant-based method to a large variety of computer vision problems and show that for most of the considered problems, the new method leads to solvers that are the same size as the the best available Grobner basis solvers and of similar accuracy. For some problems the new sparse-resultant based method leads to even smaller and more stable solvers than the state-of-the-art Grobner basis solvers. Our new method can be fully automatized and incorporated into existing tools for automatic generation of efficient polynomial solvers and as such it represents a competitive alternative to popular Grobner basis methods for minimal problems in computer vision.
****
DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection
Liming Jiang, Ren Li, Wayne Wu, Chen Qian, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2889-2898
We present our on-going effort of constructing a large- scale benchmark for face forgery detection. The first version of this benchmark, DeeperForensics-1.0, represents the largest face forgery detection dataset by far, with 60, 000 videos constituted by a total of 17.6 million frames, 10 times larger than existing datasets of the same kind. Extensive real-world perturbations are applied to obtain a more challenging benchmark of larger scale and higher diversity. All source videos in DeeperForensics-1.0 are carefully collected, and fake videos are generated by a newly proposed end-to-end face swapping framework. The quality of generated videos outperforms those in existing datasets, validated by user studies. The benchmark features a hidden test set, which contains manipulated videos achieving high deceptive scores in human evaluations. We further contribute a comprehensive study that evaluates five representative detection baselines and make a thorough analysis of different settings.

****
Shape Reconstruction by Learning Differentiable Surface Representations
Jan Bednarik, Shaifali Parashar, Erhan Gundogdu, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4716-4725
Generative models that produce point clouds have emerged as a powerful tool to represent 3D surfaces, and the best current ones rely on learning an ensemble of parametric representations. Unfortunately, they offer no control over the deformations of the surface patches that form the ensemble and thus fail to prevent them from either overlapping or collapsing into single points or lines. As a consequence, computing shape properties such as surface normals and curvatures becomes difficult and unreliable. In this paper, we show that we can exploit the inherent differentiability of deep networks to leverage differential surface properties during training so as to prevent patch collapse and strongly reduce patch overlap. Furthermore, this lets us reliably compute quantities such as surface normals and curvatures. We will demonstrate on several tasks that this yields more accurate surface reconstructions than the state-of-the-art methods in terms of normals estimation and amount of collapsed and overlapped patches.
****
Temporal Pyramid Network for Action Recognition
Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, Bolei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 591-600
Visual tempo characterizes the dynamics and the temporal scale of an action. Modeling such visual tempos of different actions facilitates their recognition. Previous works often capture the visual tempo through sampling raw videos at multiple rates and constructing an input-level frame pyramid, which usually requires a costly multi-branch network to handle. In this work we propose a generic Temporal Pyramid Network (TPN) at the feature-level, which can be flexibly integrated into 2D or 3D backbone networks in a plug-and-play manner. Two essential components of TPN, the source of features and the fusion of features, form a feature hierarchy for the backbone so that it can capture action instances at various tempos. TPN also shows consistent improvements over other challenging baselines on several action recognition datasets. Specifically, when equipped with TPN, the 3D ResNet-50 with dense sampling obtains a 2% gain on the validation set of Kinetics-400. A further analysis also reveals that TPN gains most of its improvements on action classes that have large variances in their visual tempos, validating the effectiveness of TPN.
****
Fusion-Aware Point Convolution for Online Semantic 3D Scene Segmentation
Jiazhao Zhang, Chenyang Zhu, Lintao Zheng, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4534-4543
Online semantic 3D segmentation in company with real-time RGB-D reconstruction poses special challenges such as how to perform 3D convolution directly over the progressively fused 3D geometric data, and how to smartly fuse information from frame to frame. We propose a novel fusion-aware 3D point convolution which operates directly on the geometric surface being reconstructed and exploits effectively the inter-frame correlation for high-quality 3D feature learning. This is enabled by a dedicated dynamic data structure that organizes the online acquired point cloud with local-global trees. Globally, we compile the online reconstructed 3D points into an incrementally growing coordinate interval tree, enabling fast point insertion and neighborhood query. Locally, we maintain the neighborhood information for each point using an octree whose construction benefits from the fast query of the global tree. The local octrees facilitate efficient surface-aware point convolution. Both levels of trees update dynamically and help the 3D convolution effectively exploits the temporal coherence for effective information fusion across RGB-D frames.
****
Skeleton-Based Action Recognition With Shift Graph Convolutional Network

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, Hanqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 183-192

Action recognition with skeleton data is attracting more attention in computer vision. Recently, graph convolutional networks (GCNs), which model the human body skeletons as spatiotemporal graphs, have obtained remarkable performance. However, the computational complexity of GCN-based methods are pretty heavy, typically over 15 GFLOPs for one action sample. Recent works even reach about 100 GFLOPs. Another shortcoming is that the receptive fields of both spatial graph and temporal graph are inflexible. Although some works enhance the expressiveness of spatial graph by introducing incremental adaptive modules, their performance is still limited by regular GCN structures. In this paper, we propose a novel shift graph convolutional network (Shift-GCN) to overcome both shortcomings. Instead of using heavy regular graph convolutions, our Shift-GCN is composed of novel shift graph operations and lightweight point-wise convolutions, where the shift graph operations provide flexible receptive fields for both spatial graph and temporal graph. On three datasets for skeleton-based action recognition, the proposed Shift-GCN notably exceeds the state-of-the-art methods with more than 10 times less computational complexity.

****

How Does Noise Help Robustness? Explanation and Exploration under the Neural SDE Framework

Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, Cho-Jui Hsieh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 282-290

Neural Ordinary Differential Equation (Neural ODE) has been proposed as a continuous approximation to the ResNet architecture. Some commonly used regularization mechanisms in discrete neural networks (e.g., dropout, Gaussian noise) are missing in current Neural ODE networks. In this paper, we propose a new continuous neural network framework called Neural Stochastic Differential Equation (Neural SDE), which naturally incorporates various commonly used regularization mechanisms based on random noise injection. For regularization purposes, our framework includes multiple types of noise patterns, such as dropout, additive, and multiplicative noise, which are common in plain neural networks. We provide some theoretical analyses explaining the improved robustness of our models against input perturbations. Furthermore, we demonstrate that the Neural SDE network can achieve better generalization than the Neural ODE and is more resistant to adversarial and non-adversarial input perturbations.

****

Context-Aware Group Captioning via Self-Attention and Contrastive Features

Zhuowan Li, Quan Tran, Long Mai, Zhe Lin, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3440-3450

While image captioning has progressed rapidly, existing works focus mainly on describing single images. In this paper, we introduce a new task, context-aware group captioning, which aims to describe a group of target images in the context of another group of related reference images. Context-aware group captioning requires not only summarizing information from both the target and reference image group but also contrasting between them. To solve this problem, we propose a framework combining self-attention mechanism with contrastive feature construction to effectively summarize common information from each image group while capturing discriminative information between them. To build the dataset for this task, we propose to group the images and generate the group captions based on single image captions using scene graphs matching. Our datasets are constructed on top of the public Conceptual Captions dataset and our new Stock Captions dataset. Experiments on the two datasets show the effectiveness of our method on this new task.

****

Learning to Forget for Meta-Learning

Sungyong Baik, Seokil Hong, Kyoung Mu Lee; Proceedings of the IEEE/CVF Confere

nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2379-2387
Few-shot learning is a challenging problem where the goal is to achieve generalization from only few examples. Model-agnostic meta-learning (MAML) tackles the problem by formulating prior knowledge as a common initialization across tasks, which is then used to quickly adapt to unseen tasks. However, forcibly sharing an initialization can lead to conflicts among tasks and the compromised (undesired by tasks) location on optimization landscape, thereby hindering the task adaptation. Further, we observe that the degree of conflict differs among not only tasks but also layers of a neural network. Thus, we propose task-and-layer-wise attenuation on the compromised initialization to reduce its influence. As the attenuation dynamically controls (or selectively forgets) the influence of prior knowledge for a given task and each layer, we name our method as L2F (Learn to Forget). The experimental results demonstrate that the proposed method provides faster adaptation and greatly improves the performance. Furthermore, L2F can be easily applied and improve other state-of-the-art MAML-based frameworks, illustrating its simplicity and generalizability.
****

A Self-supervised Approach for Adversarial Robustness
Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 262-271
Adversarial examples can cause catastrophic mistakes in Deep Neural Network (DNNs) based vision systems e.g., for classification, segmentation and object detection. The vulnerability of DNNs against such attacks can prove a major roadblock towards their real-world deployment. Transferability of adversarial examples demand generalizable defenses that can provide cross-task protection. Adversarial training that enhances robustness by modifying target model's parameters lacks such generalizability. On the other hand, different input processing based defenses fall short in the face of continuously evolving attacks. In this paper, we take the first step to combine the benefits of both approaches and propose a self-supervised adversarial training mechanism in the input space. By design, our defense is a generalizable approach and provides significant robustness against the unseen adversarial attacks (e.g. by reducing the success rate of translation-invariant ensemble attack from 82.6% to 31.9% in comparison to previous state-of-the-art). It can be deployed as a plug-and-play solution to protect a variety of vision systems, as we demonstrate for the case of classification, segmentation and detection.
****

Multimodal Future Localization and Emergence Prediction for Objects in Egocentric View With a Reachability Prior
Osama Makansi, Ozgun Cicek, Kevin Buchicchio, Thomas Brox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4354-4363
In this paper, we investigate the problem of anticipating future dynamics, particularly the future location of other vehicles and pedestrians, in the view of a moving vehicle. We approach two fundamental challenges: (1) the partial visibility due to the egocentric view with a single RGB camera and considerable field-of-view change due to the egomotion of the vehicle; (2) the multimodality of the distribution of future states. In contrast to many previous works, we do not assume structural knowledge from maps. We rather estimate a reachability prior for certain classes of objects from the semantic map of the present image and propagate it into the future using the planned egomotion. Experiments show that the reachability prior combined with multi-hypotheses learning improves multimodal prediction of the future location of tracked objects and, for the first time, the emergence of new objects. We also demonstrate promising zero-shot transfer to unseen datasets.
****

OccuSeg: Occupancy-Aware 3D Instance Segmentation
Lei Han, Tian Zheng, Lan Xu, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2940-2949

3D instance segmentation, with a variety of applications in robotics and augmented reality, is in large demands these days. Unlike 2D images that are projective observations of the environment, 3D models provide metric reconstruction of the scenes without occlusion or scale ambiguity. In this paper, we define "3D occupancy size", as the number of voxels occupied by each instance. It owns advantages of robustness in prediction, on which basis, OccuSeg, an occupancy-aware 3D instance segmentation scheme is proposed. Our multi-task learning produces both occupancy signal and embedding representations, where the training of spatial and feature embeddings varies with their difference in scale-aware. Our clustering scheme benefits from the reliable comparison between the predicted occupancy size and the clustered occupancy size, which encourages hard samples being correctly clustered and avoids over segmentation. The proposed approach achieves state-of-theart performance on 3 real-world datasets, i.e. ScanNetV2, S3DIS and SceneNN, while maintaining high efficiency.

****

RevealNet: Seeing Behind Objects in RGB-D Scans

Ji Hou, Angela Dai, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2098-2107

During 3D reconstruction, it is often the case that people cannot scan each individual object from all views, resulting in missing geometry in the captured scan. This missing geometry can be fundamentally limiting for many applications, e.g., a robot needs to know the unseen geometry to perform a precise grasp on an object. Thus, we introduce the task of semantic instance completion: from an incomplete RGB-D scan of a scene, we aim to detect the individual object instances and infer their complete object geometry. This will open up new possibilities for interactions with objects in a scene, for instance for virtual or robotic agents. We tackle this problem by introducing RevealNet, a new data-driven approach that jointly detects object instances and predicts their complete geometry. This enables a semantically meaningful decomposition of a scanned scene into individual, complete 3D objects, including hidden and unobserved object parts. RevealNet is an end-to-end 3D neural network architecture that leverages joint color and geometry feature learning. The fully-convolutional nature of our 3D network enables efficient inference of semantic instance completion for 3D scans at scale of large indoor environments in a single forward pass. We show that predicting complete object geometry improves both 3D detection and instance segmentation performance. We evaluate on both real and synthetic scan benchmark data for the new task, where we outperform state-of-the-art approaches by over 15 in mAP@0.5 on ScanNet, and over 18 in mAP@0.5 on SUNCG.

****

Deep Optics for Single-Shot High-Dynamic-Range Imaging

Christopher A. Metzler, Hayato Ikoma, Yifan Peng, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1375-1385

High-dynamic-range (HDR) imaging is crucial for many applications. Yet, acquiring HDR images with a single shot remains a challenging problem. Whereas modern deep learning approaches are successful at hallucinating plausible HDR content from a single low-dynamic-range (LDR) image, saturated scene details often cannot be faithfully recovered. Inspired by recent deep optical imaging approaches, we interpret this problem as jointly training an optical encoder and electronic decoder where the encoder is parameterized by the point spread function (PSF) of the lens, the bottleneck is the sensor with a limited dynamic range, and the decoder is a convolutional neural network (CNN). The lens surface is then jointly optimized with the CNN in a training phase; we fabricate this optimized optical element and attach it as a hardware add-on to a conventional camera during inference. In extensive simulations and with a physical prototype, we demonstrate that this end-to-end deep optical imaging approach to single-shot HDR imaging outperforms both purely CNN-based approaches and other PSF engineering approaches.

****

HybridPose: 6D Object Pose Estimation Under Hybrid Representations

Chen Song, Jiaru Song, Qixing Huang; Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), 2020, pp. 431-440
We introduce HybridPose, a novel 6D object pose estimation approach. HybridPose utilizes a hybrid intermediate representation to express different geometric information in the input image, including keypoints, edge vectors, and symmetry correspondences. Compared to a unitary representation, our hybrid representation allows pose regression to exploit more and diverse features when one type of predicted representation is inaccurate (e.g., because of occlusion). Different intermediate representations used by HybridPose can all be predicted by the same simple neural network, and outliers in predicted intermediate representations are filtered by a robust regression module. Compared to state-of-the-art pose estimation approaches, HybridPose is comparable in running time and is significantly more accurate. For example, on Occlusion Linemod dataset, our method achieves a prediction speed of 30 fps with a mean ADD(-S) accuracy of 79.2%, representing a 67.4% improvement from the current state-of-the-art approach.
****

Ensemble Generative Cleaning With Feedback Loops for Defending Adversarial Attacks
Jianhe Yuan, Zhihai He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 581-590
Effective defense of deep neural networks against adversarial attacks remains a challenging problem, especially under powerful white-box attacks. In this paper, we develop a new method called ensemble generative cleaning with feedback loops (EGC-FL) for effective defense of deep neural networks. The proposed EGC-FL method is based on two central ideas. First, we introduce a transformed deadzone layer into the defense network, which consists of an orthonormal transform and a deadzone-based activation function, to destroy the sophisticated noise pattern of adversarial attacks. Second, by constructing a generative cleaning network with a feedback loop, we are able to generate an ensemble of diverse estimations of the original clean image. We then learn a network to fuse this set of diverse estimations together to restore the original image. Our extensive experimental results demonstrate that our approach improves the state-of-art by large margins in both white-box and black-box attacks. It significantly improves the classification accuracy for white-box PGD attacks upon the second best method by more than 29% on the SVHN dataset and more than 39% on the challenging CIFAR-10 dataset.
****

Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification
Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3289-3299
While video-based person re-identification (Re-ID) has drawn increasing attention and made great progress in recent years, it is still very challenging to effectively overcome the occlusion problem and the visual ambiguity problem for visually similar negative samples. On the other hand, we observe that different frames of a video can provide complementary information for each other, and the structural information of pedestrians can provide extra discriminative cues for appearance features. Thus, modeling the temporal relations of different frames and the spatial relations within a frame has the potential for solving the above problems. In this work, we propose a novel Spatial-Temporal Graph Convolutional Network (STGCN) to solve these problems. The STGCN includes two GCN branches, a spatial one and a temporal one. The spatial branch extracts structural information of a human body. The temporal branch mines discriminative cues from adjacent frames. By jointly optimizing these branches, our model extracts robust spatial-temporal information that is complementary with appearance information. As shown in the experiments, our model achieves state-of-the-art results on MARS and DukeMTMC-VideoReID datasets.
****

The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks
Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, Dawn Song; Proce

edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2020, pp. 253-261
This paper studies model-inversion attacks, in which the access to a model is ab used to infer information about the training data. Since its first introduction by [??], such attacks have raised serious concerns given that training data usua lly contain privacy sensitive information. Thus far, successful model-inversion attacks have only been demonstrated on simple models, such as linear regression and logistic regression. Previous attempts to invert neural networks, even the o nes with simple architectures, have failed to produce convincing results. Here w e present a novel attack method, termed the generative model-inversion attack, w hich can invert deep neural networks with high success rates. Rather than recons tructing private training data from scratch, we leverage partial public informat ion, which can be very generic, to learn a distributional prior via generative a dversarial networks (GANs) and use it to guide the inversion process. Moreover, we theoretically prove that a model's predictive power and its vulnerability to inversion attacks are indeed two sides of the same coin---highly predictive mode ls are able to establish a strong correlation between features and labels, which coincides exactly with what an adversary exploits to mount the attacks. Our ext ensive experiments demonstrate that the proposed attack improves identification accuracy over the existing work by about 75% for reconstructing face images from a state-of-the-art face recognition classifier. We also show that differential privacy, in its canonical form, is of little avail to defend against our attacks .
****
Video Modeling With Correlation Networks
Heng Wang, Du Tran, Lorenzo Torresani, Matt Feiszli; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 352-361
Motion is a salient cue to recognize actions in video. Modern action recognition models leverage motion information either explicitly by using optical flow as i nput or implicitly by means of 3D convolutional filters that simultaneously capt ure appearance and motion information. This paper proposes an alternative approa ch based on a learnable correlation operator that can be used to establish frame -to-frame matches over convolutional feature maps in the different layers of the network. The proposed architecture enables the fusion of this explicit temporal matching information with traditional appearance cues captured by 2D convolutio n. Our correlation network compares favorably with widely-used 3D CNNs for video modeling, and achieves competitive results over the prominent two-stream networ k while being much faster to train. We empirically demonstrate that correlation networks produce strong results on a variety of video datasets, and outperform t he state of the art on four popular benchmarks for action recognition: Kinetics, Something-Something, Diving48, and Sports1M.
****
Understanding Road Layout From Videos as a Whole
Buyu Liu, Bingbing Zhuang, Samuel Schulter, Pan Ji, Manmohan Chandraker; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4414-4423
In this paper, we address the problem of inferring the layout of complex road sc enes from video sequences. To this end, we formulate it as a top-view road attri butes prediction problem and our goal is to predict these attributes for each fr ame both accurately and consistently. In contrast to prior work, we exploit the following three novel aspects: leveraging camera motions in videos, including co ntext cues and incorporating long-term video information. Specifically, we intro duce a model that aims to enforce prediction consistency in videos. Our model co nsists of one LSTM and one Feature Transform Module (FTM). The former implicitly incorporates the consistency constraint with its hidden states, and the latter explicitly takes the camera motion into consideration when aggregating informati on along videos. Moreover, we propose to incorporate context information by intr oducing road participants, e.g. objects, into our model. When the entire video s equence is available, our model is also able to encode both local and global cue

s, e.g. information from both past and future frames. Experiments on two data se ts show that: (1) Incorporating either global or contextual cues improves the pr ediction accuracy and leveraging both gives the best performance. (2) Introducin g the LSTM and FTM modules improves the prediction consistency in videos. (3) Th e proposed method outperforms the SOTA by a large margin.
****

Universal Physical Camouflage Attacks on Object Detectors
Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changq ing Zou, Ning Liu; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2020, pp. 720-729
In this paper, we study physical adversarial attacks on object detectors in the wild. Previous works mostly craft instance-dependent perturbations only for rigi d or planar objects. To this end, we propose to learn an adversarial pattern to effectively attack all instances belonging to the same object category, referred to as Universal Physical Camouflage Attack (UPC). Concretely, UPC crafts camouf lage by jointly fooling the region proposal network, as well as misleading the c lassifier and the regressor to output errors. In order to make UPC effective for non-rigid or non-planar objects, we introduce a set of transformations for mimi cking deformable properties. We additionally impose optimization constraint to m ake generated patterns look natural to human observers. To fairly evaluate the e ffectiveness of different physical-world attacks, we present the first standardi zed virtual database, AttackScenes, which simulates the real 3D world in a contr ollable and reproducible environment. Extensive experiments suggest the superior ity of our proposed UPC compared with existing physical adversarial attackers no t only in virtual environments (AttackScenes), but also in real-world physical e nvironments.
****

Ego-Topo: Environment Affordances From Egocentric Video
Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, Kristen Grauman; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2020, pp. 163-172
First-person video naturally brings the use of a physical environment to the for efront, since it shows the camera wearer interacting fluidly in a space based on his intentions. However, current methods largely separate the observed actions from the persistent space itself. We introduce a model for environment affordanc es that is learned directly from egocentric video. The main idea is to gain a hu man-centric model of a physical space (such as a kitchen) that captures (1) the primary spatial zones of interaction and (2) the likely activities they support. Our approach decomposes a space into a topological map derived from first-perso n activity, organizing an ego-video into a series of visits to the different zon es. Further, we show how to link zones across multiple related environments (e.g ., from videos of multiple kitchens) to obtain a consolidated representation of environment functionality. On EPIC-Kitchens and EGTEA+, we demonstrate our appro ach for learning scene affordances and anticipating future actions in long-form video.
****

Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector
Qi Fan, Wei Zhuo, Chi-Keung Tang, Yu-Wing Tai; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4013-4022
Conventional methods for object detection typically require a substantial amount of training data and preparing such high-quality training data is very labor-in tensive. In this paper, we propose a novel few-shot object detection network tha t aims at detecting objects of unseen categories with only a few annotated examp les. Central to our method are our Attention-RPN, Multi-Relation Detector and Co ntrastive Training strategy, which exploit the similarity between the few shot s upport set and query set to detect novel objects while suppressing false detecti on in the background. To train our network, we contribute a new dataset that con tains 1000 categories of various objects with high-quality annotations. To the b est of our knowledge, this is one of the first datasets specifically designed fo r few-shot object detection. Once our few-shot network is trained, it can detect

objects of unseen categories without further training or fine-tuning. Our method is general and has a wide range of potential applications. We produce a new state-of-the-art performance on different datasets in the few-shot setting. The dataset link is https://github.com/fanq15/Few-Shot-Object-Detection-Dataset.
****

Salience-Guided Cascaded Suppression Network for Person Re-Identification
Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3300-3310
Employing attention mechanisms to model both global and local features as a final pedestrian representation has become a trend for person re-identification (Re-ID) algorithms. A potential limitation of these methods is that they focus on the most salient features, but the re-identification of a person may rely on diverse clues masked by the most salient features in different situations, e.g., body, clothes or even shoes. To handle this limitation, we propose a novel Salience-guided Cascaded Suppression Network (SCSN) which enables the model to mine diverse salient features and integrate these features into the final representation by a cascaded manner. Our work makes the following contributions: (i) We observe that the previously learned salient features may hinder the network from learning other important information. To tackle this limitation, we introduce a cascaded suppression strategy, which enables the network to mine diverse potential useful features that be masked by the other salient features stage-by-stage and each stage integrates different feature embedding for the last discriminative pedestrian representation. (ii) We propose a Salient Feature Extraction (SFE) unit, which can suppress the salient features learned in the previous cascaded stage and then adaptively extracts other potential salient feature to obtain different clues of pedestrians. (iii) We develop an efficient feature aggregation strategy that fully increases the network's capacity for all potential salience features. Finally, experimental results demonstrate that our proposed method outperforms the state-of-the-art methods on four large-scale datasets. Especially, our approach exceeds the current best method by over 7% on the CUHK03 dataset.
****

Height and Uprightness Invariance for 3D Prediction From a Single View
Manel Baradad, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 491-500
Current state-of-the-art methods that predict 3D from single images ignore the fact that the height of objects and their upright orientation is invariant to the camera pose and intrinsic parameters. To account for this, we propose a system that directly regresses 3D world coordinates for each pixel. First, our system predicts the camera position with respect to the ground plane and its intrinsic parameters. Followed by that, it predicts the 3D position for each pixel along the rays spanned by the camera. The predicted 3D coordinates and normals are invariant to a change in the camera position or its model, and we can directly impose a regression loss on these world coordinates. Our approach yields competitive results for depth and camera pose estimation (while not being explicitly trained to predict any of these) and improves across-dataset generalization performance over existing state-of-the-art methods.
****

Projection & Probability-Driven Black-Box Attack
Jie Li, Rongrong Ji, Hong Liu, Jianzhuang Liu, Bineng Zhong, Cheng Deng, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 362-371
Generating adversarial examples in a black-box setting retains a significant challenge with vast practical application prospects. In particular, existing black-box attacks suffer from the need for excessive queries, as it is non-trivial to find an appropriate direction to optimize in the high-dimensional space. In this paper, we propose Projection & Probability-driven Black-box Attack (PPBA) to tackle this problem by reducing the solution space and providing better optimization. For reducing the solution space, we first model the adversarial perturbation optimization problem as a process of recovering frequency-sparse perturbations

with compressed sensing, under the setting that random noise in the low-frequenc y space is more likely to be adversarial. We then propose a simple method to con struct a low-frequency constrained sensing matrix, which works as a plug-and-pla y projection matrix to reduce the dimensionality. Such a sensing matrix is shown to be flexible enough to be integrated into existing methods like NES and Bandi ts_ TD . For better optimization, we perform a random walk with a probability-dr iven strategy, which utilizes all queries over the whole progress to make full u se of the sensing matrix for a less query budget. Extensive experiments show tha t our method requires at most 24% fewer queries with a higher attack success rat e compared with state-of-the-art approaches. Finally, the attack method is evalu ated on the real-world online service, i.e., Google Cloud Vision API, which furt her demonstrates our practical potentials.
****

Deep Stereo Using Adaptive Thin Volume Representation With Uncertainty Awareness
Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorth i, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2020, pp. 2524-2534
We present Uncertainty-aware Cascaded Stereo Network (UCS-Net) for 3D reconstruc tion from multiple RGB images. Multi-view stereo (MVS) aims to reconstruct fine-grained scene geometry from multi-view images. Previous learning-based MVS metho ds estimate per-view depth using plane sweep volumes (PSVs) with a fixed depth h ypothesis at each plane; this requires densely sampled planes for high accuracy, which is impractical for high-resolution depth because of limited memory. In co ntrast, we propose adaptive thin volumes (ATVs); in an ATV, the depth hypothesis of each plane is spatially varying, which adapts to the uncertainties of previo us per-pixel depth predictions. Our UCS-Net has three stages: the first stage pr ocesses a small PSV to predict low-resolution depth; two ATVs are then used in t he following stages to refine the depth with higher resolution and higher accura cy. Our ATV consists of only a small number of planes with low memory and comput ation costs; yet, it efficiently partitions local depth ranges within learned sm all uncertainty intervals. We propose to use variance-based uncertainty estimate s to adaptively construct ATVs; this differentiable process leads to reasonable and fine-grained spatial partitioning. Our multi-stage framework progressively s ub-divides the vast scene space with increasing depth resolution and precision, which enables reconstruction with high completeness and accuracy in a coarse-to-fine fashion. We demonstrate that our method achieves superior performance compa red with other learning-based MVS methods on various challenging datasets.
****

GreedyNAS: Towards Fast One-Shot NAS With Greedy Supernet
Shan You, Tao Huang, Mingmin Yang, Fei Wang, Chen Qian, Changshui Zhang; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1999-2008
Training a supernet matters for one-shot neural architecture search (NAS) method s since it serves as a basic performance estimator for different architectures (paths). Current methods mainly hold the assumption that a supernet should give a reasonable ranking over all paths. They thus treat all paths equally, and spare much effort to train paths. However, it is harsh for a single supernet to evalu ate accurately on such a huge-scale search space (e.g., 7^21). In this paper, in stead of covering all paths, we ease the burden of supernet by encouraging it to focus more on evaluation of those potentially-good ones, which are identified u sing a surrogate portion of validation data. Concretely, during training, we pro pose a multi-path sampling strategy with rejection, and greedily filter the weak paths. The training efficiency is thus boosted since the training space has bee n greedily shrunk from all paths to those potentially-good ones. Moreover, we fu rther adopt an exploration and exploitation policy by introducing an empirical c andidate path pool. Our proposed method GreedyNAS is easy-to-follow, and experim ental results on ImageNet dataset indicate that it can achieve better Top-1 accu racy under same search space and FLOPs or latency level, but with only 60% of s upernet training cost. By searching on a larger space, our GreedyNAS can also ob tain new state-of-the-art architectures.

****
Efficient Dynamic Scene Deblurring Using Spatially Variant Deconvolution Network With Optical Flow Guided Training
Yuan Yuan, Wei Su, Dandan Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3555-3564
In order to remove the non-uniform blur of images captured from dynamic scenes, many deep learning based methods design deep networks for large receptive fields and strong fitting capabilities, or use multi-scale strategy to deblur image on different scales gradually. Restricted by the fixed structures and parameters, these methods are always huge in model size to handle complex blurs. In this paper, we start from the deblurring deconvolution operation, then design an effective and real-time deblurring network. The main contributions are three folded, 1) we construct a spatially variant deconvolution network using modulated deformable convolutions, which can adjust receptive fields adaptively according to the blur features. 2) our analysis shows the sampling points of deformable convolution can be used to approximate the blur kernel, which can be simplified to bi-directional optical flows. So the position learning of sampling points can be supervised by bi-directional optical flows. 3) we build a light-weighted backbone for image restoration problem, which can balance the calculations and effectiveness well. Experimental results show that the proposed method achieves state-of-the-art deblurring performance, but with less parameters and shorter running time.
****
Learning 3D Semantic Scene Graphs From 3D Indoor Reconstructions
Johanna Wald, Helisa Dhamo, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3961-3970
Scene understanding has been of high interest in computer vision. It encompasses not only identifying objects in a scene, but also their relationships within the given context. With this goal, a recent line of works tackles 3D semantic segmentation and scene layout prediction. In our work we focus on scene graphs, a data structure that organizes the entities of a scene in a graph, where objects are nodes and their relationships modeled as edges. We leverage inference on scene graphs as a way to carry out 3D scene understanding, mapping objects and their relationships. In particular, we propose a learned method that regresses a scene graph from the point cloud of a scene. Our novel architecture is based on PointNet and Graph Convolutional Networks (GCN). In addition, we introduce 3DSSG, a semiautomatically generated dataset, that contains semantically rich scene graphs of 3D scenes. We show the application of our method in a domain-agnostic retrieval task, where graphs serve as an intermediate representation for 3D-3D and 2D-3D matching.
****
Reliable Weighted Optimal Transport for Unsupervised Domain Adaptation
Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, Jindong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4394-4403
Recently, extensive researches have been proposed to address the UDA problem, which aims to learn transferrable models for the unlabeled target domain. Among them, the optimal transport is a promising metric to align the representations of the source and target domains. However, most existing works based on optimal transport ignore the intra-domain structure, only achieving coarse pair-wise matching. The target samples distributed near the edge of the clusters, or far from their corresponding class centers are easily to be misclassified by the decision boundary learned from the source domain. In this paper, we present Reliable Weighted Optimal Transport (RWOT) for unsupervised domain adaptation, including novel Shrinking Subspace Reliability (SSR) and weighted optimal transport strategy. Specifically, SSR exploits spatial prototypical information and intra-domain structure to dynamically measure the sample-level domain discrepancy across domains. Besides, the weighted optimal transport strategy based on SSR is exploited to achieve the precise-pair-wise optimal transport procedure, which reduces negative transfer brought by the samples near decision boundaries in the target domain.

RWOT also equips with the discriminative centroid clustering exploitation strategy to learn transfer features. A thorough evaluation shows that RWOT outperforms existing state-of-the-art method on standard domain adaptation benchmarks.
****

Deblurring by Realistic Blurring
Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2737-2746
Existing deep learning methods for image deblurring typically train models using pairs of sharp images and their blurred counterparts. However, synthetically blurring images does not necessarily model the blurring process in real-world scenarios with sufficient accuracy. To address this problem, we propose a new method which combines two GAN models, i.e., a learning-to-Blur GAN (BGAN) and learning-to-DeBlur GAN (DBGAN), in order to learn a better model for image deblurring by primarily learning how to blur images. The first model, BGAN, learns how to blur sharp images with unpaired sharp and blurry image sets, and then guides the second model, DBGAN, to learn how to correctly deblur such images. In order to reduce the discrepancy between real blur and synthesized blur, a relativistic blur loss is leveraged. As an additional contribution, this paper also introduces a Real-World Blurred Image (RWBI) dataset including diverse blurry images. Our experiments show that the proposed method achieves consistently superior quantitative performance as well as higher perceptual quality on both the newly proposed dataset and the public GOPRO dataset.
****

Geometric Structure Based and Regularized Depth Estimation From 360 Indoor Imagery
Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 889-898
Motivated by the correlation between the depth and the geometric structure of a 360 indoor image, we propose a novel learning-based depth estimation framework that leverages the geometric structure of a scene to conduct depth estimation. Specifically, we represent the geometric structure of an indoor scene as a collection of corners, boundaries and planes. On the one hand, once a depth map is estimated, this geometric structure can be inferred from the estimated depth map; thus, the geometric structure functions as a regularizer for depth estimation. On the other hand, this estimation also benefits from the geometric structure of a scene estimated from an image where the structure functions as a prior. However, furniture in indoor scenes makes it challenging to infer geometric structure from depth or image data. An attention map is inferred to facilitate both depth estimation from features of the geometric structure and also geometric inferences from the estimated depth map. To validate the effectiveness of each component in our framework under controlled conditions, we render a synthetic dataset, Shanghaitech-Kujiale Indoor 360 dataset with 3550 360 indoor images. Extensive experiments on popular datasets validate the effectiveness of our solution. We also demonstrate that our method can also be applied to counterfactual depth.
****

Learning in the Frequency Domain
Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, Fengbo Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1740-1749
Deep neural networks have achieved remarkable success in computer vision tasks. Existing neural networks mainly operate in the spatial domain with fixed input sizes. For practical applications, images are usually large and have to be downsampled to the predetermined input size of neural networks. Even though the downsampling operations reduce computation and the required communication bandwidth, it removes both redundant and salient information obliviously, which results in accuracy degradation. Inspired by digital signal processing theories, we analyze the spectral bias from the frequency perspective and propose a learning-based frequency selection method to identify the trivial frequency components which can

be removed without accuracy loss. The proposed method of learning in the frequency domain leverages identical structures of the well-known neural networks, such as ResNet-50, MobileNetV2, and Mask R-CNN, while accepting the frequency-domain information as the input. Experiment results show that learning in the frequency domain with static channel selection can achieve higher accuracy than the conventional spatial downsampling approach and meanwhile further reduce the input data size. Specifically for ImageNet classification with the same input size, the proposed method achieves 1.60% and 0.63% top-1 accuracy improvements on ResNet-50 and MobileNetV2, respectively. Even with half input size, the proposed method still improves the top-1 accuracy on ResNet-50 by 1.42%. In addition, we observe a 0.8% average precision improvement on Mask R-CNN for instance segmentation on the COCO dataset.
****

BiFuse: Monocular 360 Depth Estimation via Bi-Projection Fusion
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), 2020, pp. 462-471
Depth estimation from a monocular 360 image is an emerging problem that gains popularity due to the availability of consumer-level 360 cameras and the complete surrounding sensing capability. While the standard of 360 imaging is under rapid development, we propose to predict the depth map of a monocular 360 image by mimicking both peripheral and foveal vision of the human eye. To this end, we adopt a two-branch neural network leveraging two common projections: equirectangular and cubemap projections. In particular, equirectangular projection incorporates a complete field-of-view but introduces distortion, whereas cubemap projection avoids distortion but introduces discontinuity at the boundary of the cube. Thus we propose a bi-projection fusion scheme along with learnable masks to balance the feature map from the two projections. Moreover, for the cubemap projection, we propose a spherical padding procedure which mitigates discontinuity at the boundary of each face. We apply our method to four panorama datasets and show favorable results against the existing state-of-the-art methods.
****

BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning
Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, Trevor Darrell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2636-2645
Datasets drive vision progress, yet existing driving datasets are impoverished in terms of visual content and supported tasks to study multitask learning for autonomous driving. Researchers are usually constrained to study a small set of problems on one dataset, while real-world computer vision applications require performing tasks of various complexities. We construct BDD100K, the largest driving video dataset with 100K videos and 10 tasks to evaluate the exciting progress of image recognition algorithms on autonomous driving. The dataset possesses geographic, environmental, and weather diversity, which is useful for training models that are less likely to be surprised by new conditions. Based on this diverse dataset, we build a benchmark for heterogeneous multitask learning and study how to solve the tasks together. Our experiments show that special training strategies are needed for existing models to perform such heterogeneous tasks. BDD100K opens the door for future studies in this important venue.
****

Plug-and-Play Algorithms for Large-Scale Snapshot Compressive Imaging
Xin Yuan, Yang Liu, Jinli Suo, Qionghai Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1447-1457
Snapshot compressive imaging (SCI) aims to capture the high-dimensional (usually 3D) images using a 2D sensor (detector) in a single snapshot. Though enjoying the advantages of low-bandwidth, low-power and low-cost, applying SCI to large-scale problems (HD or UHD videos) in our daily life is still challenging. The bottleneck lies in the reconstruction algorithms; they are either too slow (iterative optimization algorithms) or not flexible to the encoding process (deep learning based end-to-end networks). In this paper, we develop fast and flexible algori

thms for SCI based on the plug-and-play (PnP) framework. In addition to the wide ly used PnP-ADMM method, we further propose the PnP-GAP (generalized alternating projection) algorithm with a lower computational workload and prove the global convergence of PnP-GAP under the SCI hardware constraints. By employing deep d enoising priors, we first time show that PnP can recover a UHD color video (3840 x1644x48 with PNSR above 30dB) from a snapshot 2D measurement. Extensive results on both simulation and real datasets verify the superiority of our proposed alg orithm.

****

Single-Stage Semantic Segmentation From Image Labels

Nikita Araslanov, Stefan Roth; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2020, pp. 4253-4262

Recent years have seen a rapid growth in new approaches improving the accuracy o f semantic segmentation in a weakly supervised setting, i.e. with only image-lev el labels available for training. However, this has come at the cost of increase d model complexity and sophisticated multi-stage training procedures. This is in contrast to earlier work that used only a single stage -- training one segmenta tion network on image labels -- which was abandoned due to inferior segmentation accuracy. In this work, we first define three desirable properties of a weakly supervised method: local consistency, semantic fidelity, and completeness. Using these properties as guidelines, we then develop a segmentation-based network mo del and a self-supervised training scheme to train for semantic masks from image -level annotations in a single stage. We show that despite its simplicity, our m ethod achieves results that are competitive with significantly more complex pipe lines, substantially outperforming earlier single-stage methods.

****

Spatially-Attentive Patch-Hierarchical Network for Adaptive Motion Deblurring

Maitreya Suin, Kuldeep Purohit, A. N. Rajagopalan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3606-36 15

This paper tackles the problem of motion deblurring of dynamic scenes. Although end-to-end fully convolutional designs have recently advanced the state-of-the-a rt in non-uniform motion deblurring, their performance-complexity trade-off is s till sub-optimal. Existing approaches achieve a large receptive field by increas ing the number of generic convolution layers and kernel-size, but this comesat t he expense of of the increase in model size and inference speed. In this work, w e propose an efficient pixel adaptive and feature attentive design for handling large blur variations across different spatial locations and process each test i mage adaptively. We also propose an effective content-aware global-local filteri ng module that significantly improves performance by considering not only global dependencies but also by dynamically exploiting neighboring pixel information. We use a patch-hierarchical attentive architecture composed of the above module that implicitly discovers the spatial variations in the blur present in the inpu t image and in turn, performs local and global modulation of intermediate featur es. Extensive qualitative and quantitative comparisons with prior art on deblurr ing benchmarks demonstrate that our design offers significant improvements over the state-of-the-art in accuracy as well as speed.

****

Front2Back: Single View 3D Shape Reconstruction via Front to Back Prediction

Yuan Yao, Nico Schertler, Enrique Rosales, Helge Rhodin, Leonid Sigal, Alla Sheffer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 531-540

Reconstruction of a 3D shape from a single 2D image is a classical computer visi on problem, whose difficulty stems from the inherent ambiguity of recovering occ luded or only partially observed surfaces. Recent methods address this challenge through the use of largely unstructured neural networks that effectively distil l conditional mapping and priors over 3D shape. In this work, we induce structur e and geometric constraints by leveraging three core observations: (1) the surfa ce of most everyday objects is often almost entirely exposed from pairs of typic al opposite views; (2) everyday objects often exhibit global reflective symmetri

es which can be accurately predicted from single views; (3) opposite orthographi c views of a 3D shape share consistent silhouettes. Following these observations , we first predict orthographic 2.5D visible surface maps (depth, normal and sil houette) from perspective 2D images, and detect global reflective symmetries in this data; second, we predict the back facing depth and normal maps using as inp ut the front maps and, when available, the symmetric reflections of these maps; and finally, we reconstruct a 3D mesh from the union of these maps using a surfa ce reconstruction method best suited for this data. Our experiments demonstrate that our framework outperforms state-of-the art approaches for 3D shape reconstr uctions from 2D and 2.5D data in terms of input fidelity and details preservatio n. Specifically, we achieve 12% better performance on average in ShapeNet benchm ark dataset, and up to 19% for certain classes of objects (e.g., chairs and vess els).

****

Reverse Perspective Network for Perspective-Aware Object Counting

Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, Nicu Sebe; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4374-4383

One of the critical challenges of object counting is the dramatic scale variatio ns, which is introduced by arbitrary perspectives. We propose a reverse perspect ive network to solve the scale variations of input images, instead of generating perspective maps to smooth final outputs. The reverse perspective network expli citly evaluates the perspective distortions, and efficiently corrects the distor tions by uniformly warping the input images. Then the proposed network delivers images with similar instance scales to the regressor. Thus the regression networ k doesn't need multi-scale receptive fields to match the various scales. Besides , to further solve the scale problem of more congested areas, we enhance the cor responding regions of ground-truth with the evaluation errors. Then we force the regressor to learn from the augmented ground-truth via an adversarial process. Furthermore, to verify the proposed model, we collected a vehicle counting datas et based on Unmanned Aerial Vehicles (UAVs). The proposed dataset has fierce sca le variations. Extensive experimental results on four benchmark datasets show th e improvements of our method against the state-of-the-arts.

****

Show, Edit and Tell: A Framework for Editing Image Captions

Fawaz Sammani, Luke Melas-Kyriazi; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2020, pp. 4808-4816

Most image captioning frameworks generate captions directly from images, learnin g a mapping from visual features to natural language. However, editing existing captions can be easier than generating new ones from scratch. Intuitively, when editing captions, a model is not required to learn information that is already p resent in the caption (i.e. sentence structure), enabling it to focus on fixing details (e.g. replacing repetitive words). This paper proposes a novel approach to image captioning based on iterative adaptive refinement of an existing captio n. Specifically, our caption-editing model consisting of two sub-modules: (1) Ed itNet, a language module with an adaptive copy mechanism (Copy-LSTM) and a Selec tive Copy Memory Attention mechanism (SCMA), and (2) DCNet, an LSTM-based denois ing auto-encoder. These components enable our model to directly copy from and mo dify existing captions. Experiments demonstrate that our new approach achieves s tate of-art performance on the MS COCO dataset both with and without sequence-le vel training.

****

Self-Learning Video Rain Streak Removal: When Cyclic Consistency Meets Temporal Correspondence

Wenhan Yang, Robby T. Tan, Shiqi Wang, Jiaying Liu; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1720-1729

In this paper, we address the problem of rain streaks removal in video by develo ping a self-learned rain streak removal method, which does not require any clean groundtruth images in the training process. The method is inspired by fact that

the adjacent frames are highly correlated and can be regarded as different versions of identical scene, and rain streaks are randomly distributed along the temporal dimension. With this in mind, we construct a two-stage Self-Learned Deraining Network (SLDNet) to remove rain streaks based on both temporal correlation and consistency. In the first stage, SLDNet utilizes the temporal correlations and learns to predict the clean version of the current frame based on its adjacent rain video frames. In the second stage, SLDNet enforces the temporal consistency among different frames. It takes both the current rain frame and adjacent rain video frames to recover structural details. The first stage is responsible for reconstructing main structures, and the second stage is responsible for extracting structural details. We build our network architecture with two sub-tasks, i.e. motion estimation, and rain region detection, and optimize them jointly. Our extensive experiments demonstrate the effectiveness of our method, offering better results both quantitatively and qualitatively.
****

QEBA: Query-Efficient Boundary-Based Blackbox Attack
Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1221-1230
Machine learning (ML), especially deep neural networks (DNNs) have been widely used in various applications, including several safety-critical ones (e.g. autonomous driving). As a result, recent research about adversarial examples has raised great concerns. Such adversarial attacks can be achieved by adding a small magnitude of perturbation to the input to mislead model prediction. While several whitebox attacks have demonstrated their effectiveness, which assume that the attackers have full access to the machine learning models; blackbox attacks are more realistic in practice. In this paper, we propose a Query-Efficient Boundary-based blackbox Attack (QEBA) based only on model's final prediction labels. We theoretically show why previous boundary-based attack with gradient estimation on the whole gradient space is not efficient in terms of query numbers, and provide optimality analysis for our dimension reduction-based gradient estimation. On the other hand, we conducted extensive experiments on ImageNet and CelebA datasets to evaluate QEBA. We show that compared with the state-of-the-art blackbox attacks, QEBA is able to use a smaller number of queries to achieve a lower magnitude of perturbation with 100% attack success rate. We also show case studies of attacks on real-world APIs including MEGVII Face++ and Microsoft Azure.
****

Generating and Exploiting Probabilistic Monocular Depth Estimates
Zhihao Xia, Patrick Sullivan, Ayan Chakrabarti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 65-74
Beyond depth estimation from a single image, the monocular cue is useful in a broader range of depth inference applications and settings---such as when one can leverage other available depth cues for improved accuracy. Currently, different applications, with different inference tasks and combinations of depth cues, are solved via different specialized networks---trained separately for each application. Instead, we propose a versatile task-agnostic monocular model that outputs a probability distribution over scene depth given an input color image, as a sample approximation of outputs from a patch-wise conditional VAE. We show that this distributional output can be used to enable a variety of inference tasks in different settings, without needing to retrain for each application. Across a diverse set of applications (depth completion, user guided estimation, etc.), our common model yields results with high accuracy---comparable to or surpassing that of state-of-the-art methods dependent on application-specific networks.
****

Context-Aware Attention Network for Image-Text Retrieval
Qi Zhang, Zhen Lei, Zhaoxiang Zhang, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3536-3545
As a typical cross-modal problem, image-text bi-directional retrieval relies heavily on the joint embedding learning and similarity measure for each image-text

pair. It remains challenging because prior works seldom explore semantic corresp
ondences between modalities and semantic correlations in a single modality at th
e same time. In this work, we propose a unified Context-Aware Attention Network
(CAAN), which selectively focuses on critical local fragments (regions and words
) by aggregating the global context. Specifically, it simultaneously utilizes gl
obal inter-modal alignments and intra-modal correlations to discover latent sema
ntic relations. Considering the interactions between images and sentences in the
 retrieval process, intra-modal correlations are derived from the second-order a
ttention of region-word alignments instead of intuitively comparing the distance
 between original features. Our method achieves fairly competitive results on tw
o generic image-text retrieval datasets Flickr30K and MS-COCO.
****

From Image Collections to Point Clouds With Self-Supervised Shape and Pose Netwo
rks

K L Navaneet, Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani,
R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2020, pp. 1132-1140

Reconstructing 3D models from 2D images is one of the fundamental problems in co
mputer vision. In this work, we propose a deep learning technique for 3D object
reconstruction from a single image. Contrary to recent works that either use 3D
supervision or multi-view supervision, we use only single view images with no po
se information during training as well. This makes our approach more practical r
equiring only an image collection of an object category and the corresponding si
lhouettes. We learn both 3D point cloud reconstruction and pose estimation netwo
rks in a self-supervised manner, making use of differentiable point cloud render
er to train with 2D supervision. A key novelty of the proposed technique is to i
mpose 3D geometric reasoning into predicted 3D point clouds by rotating them wit
h randomly sampled poses and then enforcing cycle consistency on both 3D reconst
ructions and poses. In addition, using single-view supervision allows us to do t
est-time optimization on a given test image. Experiments on the synthetic ShapeN
et and real-world Pix3D datasets demonstrate that our approach, despite using le
ss supervision, can achieve competitive performance compared to pose-supervised
and multi-view supervised approaches.
****

Predicting Goal-Directed Human Attention Using Inverse Reinforcement Learning

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelin
sky, Dimitris Samaras, Minh Hoai; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2020, pp. 193-202

Human gaze behavior prediction is important for behavioral vision and for comput
er vision applications. Most models mainly focus on predicting free-viewing beha
vior using saliency maps, but do not generalize to goal-directed behavior, such
as when a person searches for a visual target object. We propose the first inver
se reinforcement learning (IRL) model to learn the internal reward function and
policy used by humans during visual search. We modeled the viewer's internal bel
ief states as dynamic contextual belief maps of object locations. These maps wer
e learned and then used to predict behavioral scanpaths for multiple target cate
gories. To train and evaluate our IRL model we created COCO-Search18, which is n
ow the largest dataset of high-quality search fixations in existence. COCO-Searc
h18 has 10 participants searching for each of 18 target-object categories in 620
2 images, making about 300,000 goal-directed fixations. When trained and evaluat
ed on COCO-Search18, the IRL model outperformed baseline models in predicting se
arch fixation scanpaths, both in terms of similarity to human search behavior an
d search efficiency. Finally, reward maps recovered by the IRL model reveal dist
inctive target-dependent patterns of object prioritization, which we interpret a
s a learned object context.
****

gDLS*: Generalized Pose-and-Scale Estimation Given Scale and Gravity Priors

Victor Fragoso, Joseph DeGol, Gang Hua; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2210-2219

Many real-world applications in augmented reality (AR), 3D mapping, and robotics

require both fast and accurate estimation of camera poses and scales from multiple images captured by multiple cameras or a single moving camera. Achieving high speed and maintaining high accuracy in a pose-and-scale estimator are often conflicting goals. To simultaneously achieve both, we exploit a priori knowledge about the solution space. We present gDLS*, a generalized-camera-model pose-and-scale estimator that utilizes rotation and scale priors. gDLS* allows an application to flexibly weigh the contribution of each prior, which is important since priors often come from noisy sensors. Compared to state-of-the-art generalized-pose-and-scale estimators (e.g., gDLS), our experiments on both synthetic and real data consistently demonstrate that gDLS* accelerates the estimation process and improves scale and pose accuracy.

****

Perceptual Quality Assessment of Smartphone Photography

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, Zhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3677-3686

As smartphones become people's primary cameras to take photos, the quality of their cameras and the associated computational photography modules has become a de facto standard in evaluating and ranking smartphones in the consumer market. We conduct so far the most comprehensive study of perceptual quality assessment of smartphone photography. We introduce the Smartphone Photography Attribute and Quality (SPAQ) database, consisting of 11,125 pictures taken by 66 smartphones, where each image is attached with so far the richest annotations. Specifically, we collect a series of human opinions for each image, including image quality, image attributes (brightness, colorfulness, contrast, noisiness, and sharpness), and scene category labels (animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others) in a well-controlled laboratory environment. The exchangeable image file format (EXIF) data for all images are also recorded to aid deeper analysis. We also make the first attempts using the database to train blind image quality assessment (BIQA) models constructed by baseline and multi-task deep neural networks. The results provide useful insights on how EXIF data, image attributes and high-level semantics interact with image quality, how next-generation BIQA models can be designed, and how better computational photography systems can be optimized on mobile devices. The database along with the proposed BIQA models are available at https://github.com/h4nwei/SPAQ.

****

When NAS Meets Robustness: In Search of Robust Architectures Against Adversarial Attacks

Minghao Guo, Yuzhe Yang, Rui Xu, Ziwei Liu, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 631-640

Recent advances in adversarial attacks uncover the intrinsic vulnerability of modern deep neural networks. Since then, extensive efforts have been devoted to enhancing the robustness of deep networks via specialized learning algorithms and loss functions. In this work, we take an architectural perspective and investigate the patterns of network architectures that are resilient to adversarial attacks. To obtain the large number of networks needed for this study, we adopt one-shot neural architecture search, training a large network for once and then finetuning the sub-networks sampled therefrom. The sampled architectures together with the accuracies they achieve provide a rich basis for our study. Our "robust architecture Odyssey" reveals several valuable observations: 1) densely connected patterns result in improved robustness; 2) under computational budget, adding convolution operations to direct connection edge is effective; 3) flow of solution procedure (FSP) matrix is a good indicator of network robustness. Based on these observations, we discover a family of robust architectures (RobNets). On various datasets, including CIFAR, SVHN, Tiny-ImageNet, and ImageNet, RobNets exhibit superior robustness performance to other widely used architectures. Notably, RobNets substantially improve the robust accuracy (5% absolute gains) under both white-box and black-box attacks, even with fewer parameter numbers. Code is available at https://github.com/gmh14/RobNets.

****
Fast Symmetric Diffeomorphic Image Registration with Convolutional Neural Networks
Tony C.W. Mok, Albert C.S. Chung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4644-4653
Diffeomorphic deformable image registration is crucial in many medical image studies, as it offers unique, special features including topology preservation and invertibility of the transformation. Recent deep learning-based deformable image registration methods achieve fast image registration by leveraging a convolutional neural network (CNN) to learn the spatial transformation from the synthetic ground truth or the similarity metric. However, these approaches often ignore the topology preservation of the transformation and the smoothness of the transformation which is enforced by a global smoothing energy function alone. Moreover, deep learning-based approaches often estimate the displacement field directly, which cannot guarantee the existence of the inverse transformation. In this paper, we present a novel, efficient unsupervised symmetric image registration method which maximizes the similarity between images within the space of diffeomorphic maps and estimates both forward and inverse transformations simultaneously. We evaluate our method on 3D image registration with a large scale brain image dataset. Our method achieves state-of-the-art registration accuracy and running time while maintaining desirable diffeomorphic properties.
****
Deep White-Balance Editing
Mahmoud Afifi, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1397-1406
We introduce a deep learning approach to realistically edit an sRGB image's white balance. Cameras capture sensor images that are rendered by their integrated signal processor (ISP) to a standard RGB (sRGB) color space encoding. The ISP rendering begins with a white-balance procedure that is used to remove the color cast of the scene's illumination. The ISP then applies a series of nonlinear color manipulations to enhance the visual quality of the final sRGB image. Recent work by [3] showed that sRGB images that were rendered with the incorrect white balance cannot be easily corrected due to the ISP's nonlinear rendering. The work in [3] proposed a k-nearest neighbor (KNN) solution based on tens of thousands of image pairs. We propose to solve this problem with a deep neural network (DNN) architecture trained in an end-to-end manner to learn the correct white balance. Our DNN maps an input image to two additional white-balance settings corresponding to indoor and outdoor illuminations. Our solution not only is more accurate than the KNN approach in terms of correcting a wrong white-balance setting but also provides the user the freedom to edit the white balance in the sRGB image to other illumination settings.
****
Convolution in the Cloud: Learning Deformable Kernels in 3D Graph Convolution Networks for Point Cloud Analysis
Zhi-Hao Lin, Sheng-Yu Huang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1800-1809
Point clouds are among the popular geometry representations for 3D vision applications. However, without regular structures like 2D images, processing and summarizing information over these unordered data points are very challenging. Although a number of previous works attempt to analyze point clouds and achieve promising performances, their performances would degrade significantly when data variations like shift and scale changes are presented. In this paper, we propose 3D Graph Convolution Networks (3D-GCN), which is designed to extract local 3D features from point clouds across scales, while shift and scale-invariance properties are introduced. The novelty of our 3D-GCN lies in the definition of learnable kernels with a graph max-pooling mechanism. We show that 3D-GCN can be applied to 3D classification and segmentation tasks, with ablation studies and visualizations verifying the design of 3D-GCN.
****

Upgrading Optical Flow to 3D Scene Flow Through Optical Expansion
Gengshan Yang, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1334-1343
We describe an approach for upgrading 2D optical flow to 3D scene flow. Our key insight is that dense optical expansion - which can be reliably inferred from monocular frame pairs - reveals changes in depth of scene elements, e.g., things moving closer will get bigger. When integrated with camera intrinsics, optical expansion can be converted into a normalized 3D scene flow vectors that provide meaningful directions of 3D movement, but not their magnitude (due to an underlying scale ambiguity). Normalized scene flow can be further "upgraded" to the true 3D scene flow knowing depth in one frame. We show that dense optical expansion between two views can be learned from annotated optical flow maps or unlabeled video sequences, and applied to a variety of dynamic 3D perception tasks including optical scene flow, LiDAR scene flow, time-to-collision estimation and depth estimation, often demonstrating significant improvement over the prior art.
****
Collaborative Distillation for Ultra-Resolution Universal Style Transfer
Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1860-1869
Universal style transfer methods typically leverage rich representations from deep Convolutional Neural Network (CNN) models (e.g., VGG-19) pre-trained on large collections of images. Despite the effectiveness, its application is heavily constrained by the large model size to handle ultra-resolution images given limited memory. In this work, we present a new knowledge distillation method (named Collaborative Distillation) for encoder-decoder based neural style transfer to reduce the convolutional filters. The main idea is underpinned by a finding that the encoder-decoder pairs construct an exclusive collaborative relationship, which is regarded as a new kind of knowledge for style transfer models. Moreover, to overcome the feature size mismatch when applying collaborative distillation, a linear embedding loss is introduced to drive the student network to learn a linear embedding of the teacher's features. Extensive experiments show the effectiveness of our method when applied to different universal style transfer approaches (WCT and AdaIN), even if the model size is reduced by 15.5 times. Especially, on WCT with the compressed models, we achieve ultra-resolution (over 40 megapixels) universal style transfer on a 12GB GPU for the first time. Further experiments on optimization-based stylization scheme show the generality of our algorithm on different stylization paradigms. Our code and trained models are available at https://github.com/mingsun-tse/collaborative-distillation.
****
L2-GCN: Layer-Wise and Learned Efficient Training of Graph Convolutional Networks
Yuning You, Tianlong Chen, Zhangyang Wang, Yang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2127-2135
Graph convolution networks (GCN) are increasingly popular in many applications, yet remain notoriously hard to train over large graph datasets. They need to compute node representations recursively from their neighbors. Current GCN training algorithms suffer from either high computational costs that grow exponentially with the number of layers, or high memory usage for loading the entire graph and node embeddings. In this paper, we propose a novel efficient layer-wise training framework for GCN (L-GCN), that disentangles feature aggregation and feature transformation during training, hence greatly reducing time and memory complexities. We present theoretical analysis for L-GCN under the graph isomorphism framework, that L-GCN leads to as powerful GCNs as the more costly conventional training algorithm does, under mild conditions. We further propose L^2-GCN, which learns a controller for each layer that can automatically adjust the training epochs per layer in L-GCN. Experiments show that L-GCN is faster than state-of-the-arts by at least an order of magnitude, with a consistent of memory usage not dependent on dataset size, while maintaining comparable prediction performance. With

the learned controller, L^2-GCN can further cut the training time in half. Our c
odes are available at https://github.com/Shen-Lab/L2-GCN.
****

Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition
Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin
Kuang, Javier Civera; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2020, pp. 2626-2635
Lifelong place recognition is an essential and challenging task in computer visi
on with vast applications in robust localization and efficient large-scale 3D re
construction. Progress is currently hindered by a lack of large, diverse, public
ly available datasets. We contribute with Mapillary Street-Level Sequences (SLS)
, a large dataset for urban and suburban place recognition from image sequences.
It contains more than 1.6 million images curated from the Mapillary collaborati
ve mapping platform. The dataset is orders of magnitude larger than current data
sources, and is designed to reflect the diversities of true lifelong learning.
It features images from 30 major cities across six continents, hundreds of disti
nct cameras, and substantially different viewpoints and capture times, spanning
all seasons over a nine year period. All images are geo-located with GPS and com
pass, and feature high-level attributes such as road type. We propose a set of b
enchmark tasks designed to push state-of-the-art performance and provide baselin
e studies. We show that current state-of-the-art methods still have a long way t
o go, and that the lack of diversity in existing datasets have prevented general
ization to new environments. The dataset and benchmarks are available for academ
ic research.
****

M-LVC: Multiple Frames Prediction for Learned Video Compression
Jianping Lin, Dong Liu, Houqiang Li, Feng Wu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3546-3554
We propose an end-to-end learned video compression scheme for low-latency scenar
ios. Previous methods are limited in using the previous one frame as reference.
Our method introduces the usage of the previous multiple frames as references. I
n our scheme, the motion vector (MV) field is calculated between the current fra
me and the previous one. With multiple reference frames and associated multiple
MV fields, our designed network can generate more accurate prediction of the cur
rent frame, yielding less residual. Multiple reference frames also help generate
MV prediction, which reduces the coding cost of MV field. We use two deep auto-
encoders to compress the residual and the MV, respectively. To compensate for th
e compression error of the auto-encoders, we further design a MV refinement netw
ork and a residual refinement network, taking use of the multiple reference fram
es as well. All the modules in our scheme are jointly optimized through a single
rate-distortion loss function. We use a step-by-step training strategy to optim
ize the entire scheme. Experimental results show that the proposed method outper
forms the existing learned video compression methods for low-latency mode. Our m
ethod also performs better than H.265 in both PSNR and MS-SSIM. Our code and mod
els are publicly available.
****

Fusing Wearable IMUs With Multi-View Images for Human Pose Estimation: A Geometr
ic Approach
Zhe Zhang, Chunyu Wang, Wenhu Qin, Wenjun Zeng; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2200-2209
We propose to estimate 3D human pose from multi-view images and a few IMUs attac
hed at person's limbs. It operates by firstly detecting 2D poses from the two si
gnals, and then lifting them to the 3D space. We present a geometric approach to
reinforce the visual features of each pair of joints based on the IMUs. This no
tably improves 2D pose estimation accuracy especially when one joint is occluded
. We call this approach Orientation Regularized Network (ORN). Then we lift the
multi-view 2D poses to the 3D space by an Orientation Regularized Pictorial Stru
cture Model (ORPSM) which jointly minimizes the projection error between the 3D
and 2D poses, along with the discrepancy between the 3D pose and IMU orientation
s. The simple two-step approach reduces the error of the state-of-the-art by a l

arge margin on a public dataset. Our code will be released at https://github.com/microsoft/imu-human-pose-estimation-pytorch.
****

DNU: Deep Non-Local Unrolling for Computational Spectral Imaging
Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1661-1671
Computational spectral imaging has been striving to capture the spectral information of the dynamic world in the last few decades. In this paper, we propose an interpretable neural network for computational spectral imaging. First, we introduce a novel data-driven prior that can adaptively exploit both the local and non-local correlations among the spectral image. Our data-driven prior is integrated as a regularizer into the reconstruction problem. Then, we propose to unroll the reconstruction problem into an optimization-inspired deep neural network. The architecture of the network has high interpretability by explicitly characterizing the image correlation and the system imaging model. Finally, we learn the complete parameters in the network through end-to-end training, enabling robust performance with high spatial-spectral fidelity. Extensive simulation and hardware experiments validate the superior performance of our method over state-of-the-art methods.
****

Single-Stage 6D Object Pose Estimation
Yinlin Hu, Pascal Fua, Wei Wang, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2930-2939
Most recent 6D pose estimation frameworks first rely on a deep network to establish correspondences between 3D object keypoints and 2D image locations and then use a variant of a RANSAC-based Perspective-n-Point (PnP) algorithm. This two-stage process, however, is suboptimal: First, it is not end-to-end trainable. Second, training the deep network relies on a surrogate loss that does not directly reflect the final 6D pose estimation task. In this work, we introduce a deep architecture that directly regresses 6D poses from correspondences. It takes as input a group of candidate correspondences for each 3D keypoint and accounts for the fact that the order of the correspondences within each group is irrelevant, while the order of the groups, that is, of the 3D keypoints, is fixed. Our architecture is generic and can thus be exploited in conjunction with existing correspondence-extraction networks so as to yield single-stage 6D pose estimation frameworks. Our experiments demonstrate that these single-stage frameworks consistently outperform their two-stage counterparts in terms of both accuracy and speed.
****

A Physics-Based Noise Formation Model for Extreme Low-Light Raw Denoising
Kaixuan Wei, Ying Fu, Jiaolong Yang, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2758-2767
Lacking rich and realistic data, learned single image denoising algorithms generalize poorly in real raw images that not resemble the data used for training. Although the problem can be alleviated by the heteroscedastic Gaussian noise model, the noise sources caused by digital camera electronics are still largely overlooked, despite their significant effect on raw measurement, especially under extremely low-light condition. To address this issue, we present a highly accurate noise formation model based on the characteristics of CMOS photosensors, thereby enabling us to synthesize realistic samples that better match the physics of image formation process. Given the proposed noise model, we additionally propose a method to calibrate the noise parameters for available modern digital cameras, which is simple and reproducible for any new device. We systematically study the generalizability of a neural network trained with existing schemes, by introducing a new low-light denoising dataset that covers many modern digital cameras from diverse brands. Extensive empirical results collectively show that by utilizing our proposed noise formation model, a network can reach the capability as if it had been trained with rich real data, which demonstrates the effectiveness of our noise formation model.

****
AdaBits: Neural Network Quantization With Adaptive Bit-Widths
Qing Jin, Linjie Yang, Zhenyu Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2146-2156
Deep neural networks with adaptive configurations have gained increasing attention due to the instant and flexible deployment of these models on platforms with different resource budgets. In this paper, we investigate a novel option to achieve this goal by enabling adaptive bit-widths of weights and activations in the model. We first examine the benefits and challenges of training quantized model with adaptive bit-widths, and then experiment with several approaches including direct adaptation, progressive training and joint training. We discover that joint training is able to produce comparable performance on the adaptive model as individual models. We also propose a new technique named Switchable Clipping Level (S-CL) to further improve quantized models at the lowest bit-width. With our proposed techniques applied on a bunch of models including MobileNet V1/V2 and ResNet50, we demonstrate that bit-width of weights and activations is a new option for adaptively executable deep neural networks, offering a distinct opportunity for improved accuracy-efficiency trade-off as well as instant adaptation according to the platform constraints in real-world applications.
****
A Lighting-Invariant Point Processor for Shading
Kathryn Heal, Jialiang Wang, Steven J. Gortler, Todd Zickler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 94-102
Under the conventional diffuse shading model with unknown directional lighting, the set of quadratic surface shapes that are consistent with the spatial derivatives of intensity at a single image point is a two-dimensional algebraic variety embedded in the five-dimensional space of quadratic shapes. We describe the geometry of this variety, and we introduce a concise feedforward model that computes an explicit, differentiable approximation of the variety from the intensity and its derivatives at any single image point. The result is a parallelizable processor that operates at each image point and produces a lighting-invariant descriptor of the continuous set of compatible surface shapes at the point. We describe two applications of this processor: two-shot uncalibrated photometric stereo and quadratic-surface shape from shading.
****
Learning a Reinforced Agent for Flexible Exposure Bracketing Selection
Zhouxia Wang, Jiawei Zhang, Mude Lin, Jiong Wang, Ping Luo, Jimmy Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1820-1828
Automatically selecting exposure bracketing (images exposed differently) is important to obtain a high dynamic range image by using multi-exposure fusion. Unlike previous methods that have many restrictions such as requiring camera response function, sensor noise model, and a stream of preview images with different exposures (not accessible in some scenarios e.g. mobile applications), we propose a novel deep neural network to automatically select exposure bracketing, named EBSNet, which is sufficiently flexible without having the above restrictions. EBSNet is formulated as a reinforced agent that is trained by maximizing rewards provided by a multi-exposure fusion network (MEFNet). By utilizing the illumination and semantic information extracted from just a single auto-exposure preview image, EBSNet enables to select an optimal exposure bracketing for multi-exposure fusion. EBSNet and MEFNet can be jointly trained to produce favorable results against recent state-of-the-art approaches. To facilitate future research, we provide a new benchmark dataset for multi-exposure selection and fusion.
****
Focus on Defocus: Bridging the Synthetic to Real Domain Gap for Depth Estimation
Maxim Maximov, Kevin Galim, Laura Leal-Taixe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1071-1080
Data-driven depth estimation methods struggle with the generalization outside their training scenes due to the immense variability of the real-world scenes. Thi

s problem can be partially addressed by utilising synthetically generated images , but closing the synthetic-real domain gap is far from trivial. In this paper, we tackle this issue by using domain invariant defocus blur as direct supervisio n. We leverage defocus cues by using a permutation invariant convolutional neura l network that encourages the network to learn from the differences between imag es with a different point of focus. Our proposed network uses the defocus map as an intermediate supervisory signal. We are able to train our model completely o n synthetic data and directly apply it to a wide range of real-world images. We evaluate our model on synthetic and real datasets, showing compelling generaliza tion results and state-of-the-art depth prediction. The dataset and code are ava ilable at https://github.com/dvl-tum/defocus-net.

****

Defending Against Universal Attacks Through Selective Feature Regeneration

Tejas Borkar, Felix Heide, Lina Karam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 709-719

Deep neural network (DNN) predictions have been shown to be vulnerable to carefu lly crafted adversarial perturbations. Specifically, image-agnostic (universal a dversarial) perturbations added to any image can fool a target network into maki ng erroneous predictions. Departing from existing defense strategies that work m ostly in the image domain, we present a novel defense which operates in the DNN feature domain and effectively defends against such universal perturbations. Our approach identifies pre-trained convolutional features that are most vulnerable to adversarial noise and deploys trainable feature regeneration units which tra nsform these DNN filter activations into resilient features that are robust to u niversal perturbations. Regenerating only the top 50% adversarially susceptible activations in at most 6 DNN layers and leaving all remaining DNN activations un changed, we outperform existing defense strategies across different network arch itectures by more than 10% in restored accuracy. We show that without any additi onal modification, our defense trained on ImageNet with one type of universal at tack examples effectively defends against other types of unseen universal attack s.

****

Cross-View Correspondence Reasoning Based on Bipartite Graph Convolutional Netwo rk for Mammogram Mass Detection

Yuhang Liu, Fandong Zhang, Qianyi Zhang, Siwen Wang, Yizhou Wang, Yizhou Yu ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2020, pp. 3812-3822

Mammogram mass detection is of great clinical significance due to its high propo rtion in breast cancers. The information from cross views (i.e., mediolateral ob lique and cranio-caudal) is highly related and complementary, and is helpful to make comprehensive decisions. However, unlike radiologists who are able to recog nize masses with reasoning ability in cross-view images, most existing methods l ack the ability to reason under the guidance of domain knowledge, thus it limits the performance. In this paper, we introduce bipartite graph convolutional netw ork to endow existing methods with cross-view reasoning ability of radiologists in mammogram mass detection. The bipartite node sets are constructed by cross-vi ew images respectively to represent relatively consistent regions in breasts, wh ile the bipartite edge learns to model both inherent cross-view geometric constr aints and appearance similarities between correspondences. Based on the bipartit e graph, the information propagates methodically through correspondences and ena bles spatial visual features equipped with customized cross-view reasoning abili ty. Experimental results on DDSM dataset demonstrate that the proposed algorithm achieves state-of-the-art performance. Besides, visual analysis shows the model has a clear physical meaning, which is helpful for radiologists in clinical int erpretation.

****

Image Search With Text Feedback by Visiolinguistic Attention Learning

Yanbei Chen, Shaogang Gong, Loris Bazzani; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3001-3011

Image search with text feedback has promising impacts in various real-world appl

ications, such as e-commerce and internet search. Given a reference image and te
xt feedback from user, the goal is to retrieve images that not only resemble the
 input image, but also change certain aspects in accordance with the given text.
 This is a challenging task as it requires the synergistic understanding of both
 image and text. In this work, we tackle this task by a novel Visiolinguistic At
tention Learning (VAL) framework. Specifically, we propose a composite transform
er that can be seamlessly plugged in a CNN to selectively preserve and transform
 the visual features conditioned on language semantics. By inserting multiple co
mposite transformers at varying depths, VAL is incentive to encapsulate the mult
i-granular visiolinguistic information, thus yielding an expressive representati
on for effective image search. We conduct comprehensive evaluation on three data
sets: Fashion200k, Shoes and FashionIQ. Extensive experiments show our model exc
eeds existing approaches on all datasets, demonstrating consistent superiority i
n coping with various text feedbacks, including attribute-like and natural langu
age descriptions.
****

Joint 3D Instance Segmentation and Object Detection for Autonomous Driving
Dingfu Zhou, Jin Fang, Xibin Song, Liu Liu, Junbo Yin, Yuchao Dai, Hongdon
g Li, Ruigang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2020, pp. 1839-1849
Currently, in Autonomous Driving (AD), most of the 3D object detection framework
s (either anchor- or anchor-free-based) consider the detection as a Bounding Box
 (BBox) regression problem. However, this compact representation is not sufficie
nt to explore all the information of the objects. To tackle this problem, we pro
pose a simple but practical detection framework to jointly predict the 3D BBox a
nd instance segmentation. For instance segmentation, we propose a Spatial Embedd
ings (SEs) strategy to assemble all foreground points into their corresponding o
bject centers. Base on the SE results, the object proposals can be generated bas
ed on a simple clustering strategy. For each cluster, only one proposal is gener
ated. Therefore, the Non-Maximum Suppression (NMS) process is no longer needed h
ere. Finally, with our proposed instance-aware ROI pooling, the BBox is refined
by a second-stage network. Experimental results on the public KITTI dataset show
 that the proposed SEs can significantly improve the instance segmentation resul
ts compared with other feature embedding-based method. Meanwhile, it also outper
forms most of the 3D object detectors on the KITTI testing benchmark.
****

DUNIT: Detection-Based Unsupervised Image-to-Image Translation
Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, Mathieu Salzmann; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2020, pp. 4787-4796
Image-to-image translation has made great strides in recent years, with current
techniques being able to handle unpaired training images and to account for the
multi-modality of the translation problem. Despite this, most methods treat the
image as a whole, which makes the results they produce for content-rich scenes l
ess realistic. In this paper, we introduce a Detection-based Unsupervised Image-
to-image Translation (DUNIT) approach that explicitly accounts for the object in
stances in the translation process. To this end, we extract separate representat
ions for the global image and for the instances, which we then fuse into a commo
n representation from which we generate the translated image. This allows us to
preserve the detailed content of object instances, while still modeling the fact
 that we aim to produce an image of a single consistent scene. We introduce an i
nstance consistency loss to maintain the coherence between the detections. Furth
ermore, by incorporating a detector into our architecture, we can still exploit
object instances at test time. As evidenced by our experiments, this allows us t
o outperform the state-of-the-art unsupervised image-to-image translation method
s. Furthermore, our approach can also be used as an unsupervised domain adaptati
on strategy for object detection, and it also achieves state-of-the-art performa
nce on this task.
****

Self-Supervised Human Depth Estimation From Monocular Videos

Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, Ping Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 650-659

Previous methods on estimating detailed human depth often require supervised training with 'ground truth' depth data. This paper presents a self-supervised method that can be trained on YouTube videos without known depth, which makes training data collection simple and improves the generalization of the learned network. The self-supervised learning is achieved by minimizing a photo-consistency loss, which is evaluated between a video frame and its neighboring frames warped according to the estimated depth and the 3D non-rigid motion of the human body. To solve this non-rigid motion, we first estimate a rough SMPL model at each video frame and compute the non-rigid body motion accordingly, which enables self-supervised learning on estimating the shape details. Experiments demonstrate that our method enjoys better generalization, and performs much better on data in the wild.

****

Enhancing Intrinsic Adversarial Robustness via Feature Pyramid Decoder

Guanlin Li, Shuya Ding, Jun Luo, Chang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 800-808

Whereas adversarial training is employed as the main defence strategy against specific adversarial samples, it has limited generalization capability and incurs excessive time complexity. In this paper, we propose an attack-agnostic defence framework to enhance the intrinsic robustness of neural networks, without jeopardizing the ability of generalizing clean samples. Our Feature Pyramid Decoder (FPD) framework applies to all block-based convolutional neural networks (CNNs). It implants denoising and image restoration modules into a targeted CNN, and it also constraints the Lipschitz constant of the classification layer. Moreover, we propose a two-phase strategy to train the FPD-enhanced CNN, utilizing e-neighborhood noisy images with multi-task and self-supervised learning. Evaluated against a variety of white-box and black-box attacks, we demonstrate that FPD-enhanced CNNs gain sufficient robustness against general adversarial samples on MNIST, SVHN and CALTECH. In addition, if we further conduct adversarial training, the FPD-enhanced CNNs perform better than their non-enhanced versions.

****

Two-Shot Spatially-Varying BRDF and Shape Estimation

Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3982-3991

Capturing the shape and spatially-varying appearance (SVBRDF) of an object from images is a challenging task that has applications in both computer vision and graphics. Traditional optimization-based approaches often need a large number of images taken from multiple views in a controlled environment. Newer deep learning-based approaches require only a few input images, but the reconstruction quality is not on par with optimization techniques. We propose a novel deep learning architecture with a stage-wise estimation of shape and SVBRDF. The previous predictions guide each estimation, and a joint refinement network later refines both SVBRDF and shape. We follow a practical mobile image capture setting and use unaligned two-shot flash and no-flash images as input. Both our two-shot image capture and network inference can run on mobile hardware. We also create a large-scale synthetic training dataset with domain-randomized geometry and realistic materials. Extensive experiments on both synthetic and real-world datasets show that our network trained on a synthetic dataset can generalize well to real-world images. Comparisons with recent approaches demonstrate the superior performance of the proposed approach.

****

Polarized Non-Line-of-Sight Imaging

Kenichiro Tanaka, Yasuhiro Mukaigawa, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2136-2145

This paper presents a method of passive non-line-of-sight (NLOS) imaging using p

olarization cues. A key observation is that the oblique light has a different po larimetric signal. It turns out this effect is due to the polarization axis rota tion, a phenomena which can be used to better condition the light transport matr ix for non-line-of-sight imaging. Our analysis and results show that the use of a polarization for NLOS is both a standalone technique, as well as an enhancemen t technique to boost the results of other forms of passive NLOS imaging. We make a surprising finding that, despite 50% light attenuation from polarization opti cs, the gains from polarized NLOS are overall superior to unpolarized NLOS.
****

End-to-End Optimization of Scene Layout
Andrew Luo, Zhoutong Zhang, Jiajun Wu, Joshua B. Tenenbaum; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, p p. 3754-3763
We propose an end-to-end variational generative model for scene layout synthesis conditioned on scene graphs. Unlike unconditional scene layout generation, we u se scene graphs as an abstract but general representation to guide the synthesis of diverse scene layouts that satisfy relationships included in the scene graph . This gives rise to more flexible control over the synthesis process, allowing various forms of inputs such as scene layouts extracted from sentences or inferr ed from a single color image. Using our conditional layout synthesizer, we can g enerate various layouts that share the same structure of the input example. In a ddition to this conditional generation design, we also integrate a differentiabl e rendering module that enables layout refinement using only 2D projections of t he scene. Given a depth and a semantics map, the differentiable rendering module enables optimizing over the synthesized layout to fit the given input in an ana lysis-by-synthesis fashion. Experiments suggest that our model achieves higher a ccuracy and diversity in conditional scene synthesis and allows exemplar-based s cene generation from various input forms.
****

HVNet: Hybrid Voxel Network for LiDAR Based 3D Object Detection
Maosheng Ye, Shuangjie Xu, Tongyi Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1631-1640
We present Hybrid Voxel Network (HVNet), a novel one-stage unified network for p oint cloud based 3D object detection for autonomous driving. Recent studies show that 2D voxelization with per voxel PointNet style feature extractor leads to a ccurate and efficient detector for large 3D scenes. Since the size of the featur e map determines the computation and memory cost, the size of the voxel becomes a parameter that is hard to balance. A smaller voxel size gives a better perform ance, especially for small objects, but a longer inference time. A larger voxel can cover the same area with a smaller feature map, but fails to capture intrica te features and accurate location for smaller objects. We present a Hybrid Voxel network that solves this problem by fusing voxel feature encoder (VFE) of diffe rent scales at point-wise level and project into multiple pseudo-image feature m aps. We further propose an attentive voxel feature encoding that outperforms pla in VFE and a feature fusion pyramid network to aggregate multi-scale information at feature map level. Experiments on the KITTI benchmark show that a single HVN et achieves the best mAP among all existing methods with a real time inference s peed of 31Hz.
****

Action Modifiers: Learning From Adverbs in Instructional Videos
Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 , pp. 868-878
We present a method to learn a representation for adverbs from instructional vid eos using weak supervision from the accompanying narrations. Key to our method i s the fact that the visual representation of the adverb is highly dependent on t he action to which it applies, although the same adverb will modify multiple act ions in a similar way. For instance, while 'spread quickly' and 'mix quickly' wi ll look dissimilar, we can learn a common representation that allows us to recog nize both, among other actions. We formulate this as an embedding problem, and u

se scaled dot product attention to learn from weakly-supervised video narrations . We jointly learn adverbs as invertible transformations which operate on the em bedding space, so as to add or remove the effect of the adverb. As there is no p rior work on weakly supervised learning from adverbs, we gather paired action-ad verb annotations from a subset of the HowTo100M dataset, for 6 adverbs: quickly/ slowly, finely/coarsely and partially/completely. Our method outperforms all bas elines for video-to-adverb retrieval with a performance of 0.719 mAP. We also de monstrate our model's ability to attend to the relevant video parts in order to determine the adverb for a given action.
****

Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement
Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong , Runmin Cong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2020, pp. 1780-1789
The paper presents a novel method, Zero-Reference Deep Curve Estimation (Zero-DC E), which formulates light enhancement as a task of image-specific curve estimat ion with a deep network. Our method trains a lightweight deep network, DCE-Net, to estimate pixel-wise and high-order curves for dynamic range adjustment of a g iven image. The curve estimation is specially designed, considering pixel value range, monotonicity, and differentiability. Zero-DCE is appealing in its relaxed assumption on reference images, i.e., it does not require any paired or unpaire d data during training. This is achieved through a set of carefully formulated n on-reference loss functions, which implicitly measure the enhancement quality an d drive the learning of the network. Our method is efficient as image enhancemen t can be achieved by an intuitive and simple nonlinear curve mapping. Despite it s simplicity, we show that it generalizes well to diverse lighting conditions. E xtensive experiments on various benchmarks demonstrate the advantages of our met hod over state-of-the-art methods qualitatively and quantitatively. Furthermore, the potential benefits of our Zero-DCE to face detection in the dark are discus sed.
****

FPConv: Learning Local Flattening for Point Convolution
Yiqun Lin, Zizheng Yan, Haibin Huang, Dong Du, Ligang Liu, Shuguang Cui, X iaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2020, pp. 4293-4302
We introduce FPConv, a novel surface-style convolution operator designed for 3D point cloud analysis. Unlike previous methods, FPConv doesn't require transformi ng to intermediate representation like 3D grid or graph and directly works on su rface geometry of point cloud. To be more specific, for each point, FPConv perfo rms a local flattening by automatically learning a weight map to softly project surrounding points onto a 2D grid. Regular 2D convolution can thus be applied fo r efficient feature learning. FPConv can be easily integrated into various netwo rk architectures for tasks like 3D object classification and 3D scene segmentati on, and achieve comparable performance with existing volumetric-type convolution s. More importantly, our experiments also show that FPConv can be a complementar y of volumetric convolutions and jointly training them can further boost overall performance into state-of-the-art results.
****

View-GCN: View-Based Graph Convolutional Network for 3D Shape Analysis
Xin Wei, Ruixuan Yu, Jian Sun; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2020, pp. 1850-1859
View-based approach that recognizes 3D shape through its projected 2D images has achieved state-of-the-art results for 3D shape recognition. The major challenge for view-based approach is how to aggregate multi-view features to be a global shape descriptor. In this work, we propose a novel view-based Graph Convolutiona l Neural Network, dubbed as view-GCN, to recognize 3D shape based on graph repre sentation of multiple views in flexible view configurations. We first construct view-graph with multiple views as graph nodes, then design a graph convolutional neural network over view-graph to hierarchically learn discriminative shape des criptor considering relations of multiple views. The view-GCN is a hierarchical

network based on local and non-local graph convolution for feature transform, an d selective view-sampling for graph coarsening. Extensive experiments on benchma rk datasets show that view-GCN achieves state-of-the-art results for 3D shape cl assification and retrieval.
****

Footprints and Free Space From a Single Color Image
Jamie Watson, Michael Firman, Aron Monszpart, Gabriel J. Brostow; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11-20
Understanding the shape of a scene from a single color image is a formidable com puter vision task. However, most methods aim to predict the geometry of surfaces that are visible to the camera, which is of limited use when planning paths for robots or augmented reality agents. Such agents can only move when grounded on a traversable surface, which we define as the set of classes which humans can al so walk over, such as grass, footpaths and pavement. Models which predict beyond the line of sight often parameterize the scene with voxels or meshes, which can be expensive to use in machine learning frameworks. We introduce a model to pre dict the geometry of both visible and occluded traversable surfaces, given a sin gle RGB image as input. We learn from stereo video sequences, using camera poses , per-frame depth and semantic segmentation to form training data, which is used to supervise an image-to-image network. We train models from the KITTI driving dataset, the indoor Matterport dataset, and from our own casually captured stere o footage. We find that a surprisingly low bar for spatial coverage of training scenes is required. We validate our algorithm against a range of strong baseline s, and include an assessment of our predictions for a path-planning task.
****

Towards Efficient Model Compression via Learned Global Ranking
Ting-Wu Chin, Ruizhou Ding, Cha Zhang, Diana Marculescu; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1518-1528
Pruning convolutional filters has demonstrated its effectiveness in compressing ConvNets. Prior art in filter pruning requires users to specify a target model c omplexity (e.g., model size or FLOP count) for the resulting architecture. Howev er, determining a target model complexity can be difficult for optimizing variou s embodied AI applications such as autonomous robots, drones, and user-facing ap plications. First, both the accuracy and the speed of ConvNets can affect the pe rformance of the application. Second, the performance of the application can be hard to assess without evaluating ConvNets during inference. As a consequence, f inding a sweet-spot between the accuracy and speed via filter pruning, which nee ds to be done in a trial-and-error fashion, can be time-consuming. This work tak es a first step toward making this process more efficient by altering the goal o f model compression to producing a set of ConvNets with various accuracy and lat ency trade-offs instead of producing one ConvNet targeting some pre-defined late ncy constraint. To this end, we propose to learn a global ranking of the filters across different layers of the ConvNet, which is used to obtain a set of ConvNe t architectures that have different accuracy/latency trade-offs by pruning the b ottom-ranked filters. Our proposed algorithm, LeGR, is shown to be 2x to 3x fast er than prior work while having comparable or better performance when targeting seven pruned ResNet-56 with different accuracy/FLOPs profiles on the CIFAR-100 d ataset. Additionally, we have evaluated LeGR on ImageNet and Bird-200 with ResNe t-50 and Mo- bileNetV2 to demonstrate its effectiveness. Code available at https ://github.com/cmu-enyac/LeGR.
****

Learning Multiview 3D Point Cloud Registration
Zan Gojcic, Caifa Zhou, Jan D. Wegner, Leonidas J. Guibas, Tolga Birdal; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1759-1769
We present a novel, end-to-end learnable, multiview 3D point cloud registration algorithm. Registration of multiple scans typically follows a two-stage pipeline : the initial pairwise alignment and the globally consistent refinement. The for

mer is often ambiguous due to the low overlap of neighboring point clouds, symmetries and repetitive scene parts. Therefore, the latter global refinement aims at establishing the cyclic consistency across multiple scans and helps in resolving the ambiguous cases. In this paper we propose, to the best of our knowledge, the first end-to-end algorithm for joint learning of both parts of this two-stage problem. Experimental evaluation on well accepted benchmark datasets shows that our approach outperforms the state-of-the-art by a significant margin, while being end-to-end trainable and computationally less costly. Moreover, we present detailed analysis and an ablation study that validate the novel components of our approach. The source code and pretrained models are publicly available under https://github.com/zgojcic/3D_multiview_reg.

****

Transformation GAN for Unsupervised Image Synthesis and Representation Learning
Jiayu Wang, Wengang Zhou, Guo-Jun Qi, Zhongqian Fu, Qi Tian, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 472-481

Generative Adversarial Networks (GAN) have shown promising performance in image synthesis and unsupervised learning (USL). In most cases, however, the representations extracted from unsupervised GAN are usually unsatisfactory in other computer vision tasks. By using conditional GAN (CGAN), this problem could be solved to some extent, but the main drawback of such models is the necessity for labeled data. To improve both image synthesis quality and representation learning performance under the unsupervised setting, in this paper, we propose a simple yet effective Transformation Generative Adversarial Networks (TrGAN). In our approach, instead of capturing the joint distribution of image-label pairs p(x,y) as in conditional GAN, we try to estimate the joint distribution of transformed image t(x) and transformation t. Specifically, given a randomly sampled transformation t, we train the discriminator to give an estimate of input transformation, while following the adversarial training scheme of the original GAN. In addition, intermediate feature matching as well as feature-transform matching methods are introduced to strengthen the regularization on the generated features. To evaluate the quality of both generated samples and extracted representations, extensive experiments are conducted on four public datasets. The experimental results on the quality of both the synthesized images and the extracted representations demonstrate the effectiveness of our method.

****

Robustness Guarantees for Deep Neural Networks on Videos
Min Wu, Marta Kwiatkowska; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 311-320

The widespread adoption of deep learning models places demands on their robustness. In this paper, we consider the robustness of deep neural networks on videos, which comprise both the spatial features of individual frames extracted by a convolutional neural network and the temporal dynamics between adjacent frames captured by a recurrent neural network. To measure robustness, we study the maximum safe radius problem, which computes the minimum distance from the optical flow sequence obtained from a given input to that of an adversarial example in the neighbourhood of the input. We demonstrate that, under the assumption of Lipschitz continuity, the problem can be approximated using finite optimisation via discretising the optical flow space, and the approximation has provable guarantees. We then show that the finite optimisation problem can be solved by utilising a two-player turn-based game in a cooperative setting, where the first player selects the optical flows and the second player determines the dimensions to be manipulated in the chosen flow. We employ an anytime approach to solve the game, in the sense of approximating the value of the game by monotonically improving its upper and lower bounds. We exploit a gradient-based search algorithm to compute the upper bounds, and the admissible A* algorithm to update the lower bounds. Finally, we evaluate our framework on the UCF101 video dataset.

****

Neural Data Server: A Large-Scale Search Engine for Transfer Learning Data
Xi Yan, David Acuna, Sanja Fidler; Proceedings of the IEEE/CVF Conference on C

omputer Vision and Pattern Recognition (CVPR), 2020, pp. 3893-3902
Transfer learning has proven to be a successful technique to train deep learning models in the domains where little training data is available. The dominant approach is to pretrain a model on a large generic dataset such as ImageNet and finetune its weights on the target domain. However, in the new era of an ever-increasing number of massive datasets, selecting the relevant data for pretraining is a critical issue. We introduce Neural Data Server (NDS), a large-scale search engine for finding the most useful transfer learning data to the target domain. NDS consists of a dataserver which indexes several large popular image datasets, and aims to recommend data to a client, an end-user with a target application with its own small labeled dataset. The dataserver represents large datasets with a much more compact mixture-of-experts model, and employs it to perform data search in a series of dataserver-client transactions at a low computational cost. We show the effectiveness of NDS in various transfer learning scenarios, demonstrating state-of-the-art performance on several target datasets and tasks such as image classification, object detection and instance segmentation. Neural Data Server is available as a web-service at http://aidemo s.cs.toronto.edu/nds/.
****

Instance Guided Proposal Network for Person Search
Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2585-2594
Person detection networks have been widely used in person search. These detectors discriminate persons from the background and generate proposals of all the persons from a gallery of scene images for each query. However, such a large number of proposals have a negative influence on the following identity matching process because many distractors are involved. In this paper, we propose a new detection network for person search, named Instance Guided Proposal Network (IGPN), which can learn the similarity between query persons and proposals. Thus, we can decrease proposals according to the similarity scores. To incorporate information of the query into the detection network, we introduce the Siamese region proposal network to Faster-RCNN and we propose improved cross-correlation layers to alleviate the imbalance of parameters distribution. Furthermore, we design a local relation block and a global relation branch to leverage the proposal-proposal relations and query-scene relations, respectively. Extensive experiments show that our method improves the person search performance through decreasing proposals and achieves competitive performance on two large person search benchmark datasets, CUHK-SYSU and PRW.
****

Deep Representation Learning on Long-Tailed Data: A Learnable Embedding Augmentation Perspective
Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, Wenhui Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2970-2979
This paper considers learning deep features from long-tailed data. We observe that in the deep feature space, the head classes and the tail classes present different distribution patterns. The head classes have a relatively large spatial span, while the tail classes have a significantly small spatial span, due to the lack of intra-class diversity. This uneven distribution between head and tail classes distorts the overall feature space, which compromises the discriminative ability of the learned features. In response, we seek to expand the distribution of the tail classes during training, so as to alleviate the distortion of the feature space. To this end, we propose to augment each instance of the tail classes with certain disturbances in the deep feature space. With the augmentation, a specified feature vector becomes a set of probable features scattered around itself, which is analogical to an atomic nucleus surrounded by the electron cloud. Intuitively, we name it as "feature cloud". The intra-class distribution of the feature cloud is learned from the head classes, and thus provides higher intra-class variation to the tail classes. Consequentially, it alleviates the distortion of the learned feature space, and improves deep representation learning on long

tailed data. Extensive experimental evaluations on person re-identification and face recognition tasks confirm the effectiveness of our method.
****

Non-Line-of-Sight Surface Reconstruction Using the Directional Light-Cone Transform
Sean I. Young, David B. Lindell, Bernd Girod, David Taubman, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1407-1416
We propose a joint albedo-normal approach to non-line-of-sight (NLOS) surface reconstruction using the directional light-cone transform (D-LCT). While current NLOS imaging methods reconstruct either the albedo or surface normals of the hidden scene, the two quantities provide complementary information of the scene, so an efficient method to estimate both simultaneously is desirable. We formulate the recovery of the two quantities as a vector deconvolution problem, and solve it via Cholesky-Wiener decomposition. We demonstrate that surfaces fitted non-parametrically using our recovered normals are more accurate than those produced with NLOS surface reconstruction methods recently proposed, and are 1,000 times faster to compute than using inverse rendering.
****

IntrA: 3D Intracranial Aneurysm Dataset for Deep Learning
Xi Yang, Ding Xia, Taichi Kin, Takeo Igarashi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2656-2666
Medicine is an important application area for deep learning models. Research in this field is a combination of medical expertise and data science knowledge. In this paper, instead of 2D medical images, we introduce an open-access 3D intracranial aneurysm dataset, IntrA, that makes the application of points-based and mesh-based classification and segmentation models available. Our dataset can be used to diagnose intracranial aneurysms and to extract the neck for a clipping operation in medicine and other areas of deep learning, such as normal estimation and surface reconstruction. We provide a large-scale benchmark of classification and part segmentation by testing state-of-the-art networks. We also discuss the performance of each method and demonstrate the challenges of our dataset. The published dataset can be accessed here: https://github.com/intra2d2019/IntrA.
****

3D-ZeF: A 3D Zebrafish Tracking Benchmark Dataset
Malte Pedersen, Joakim Bruslund Haurum, Stefan Hein Bengtson, Thomas B. Moeslund; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2426-2436
In this work we present a novel publicly available stereo based 3D RGB dataset for multi-object zebrafish tracking, called 3D-ZeF. Zebrafish is an increasingly popular model organism used for studying neurological disorders, drug addiction, and more. Behavioral analysis is often a critical part of such research. However, visual similarity, occlusion, and erratic movement of the zebrafish makes robust 3D tracking a challenging and unsolved problem. The proposed dataset consists of eight sequences with a duration between 15-120 seconds and 1-10 free moving zebrafish. The videos have been annotated with a total of 86,400 points and bounding boxes. Furthermore, we present a complexity score and a novel open-source modular baseline system for 3D tracking of zebrafish. The performance of the system is measured with respect to two detectors: a naive approach and a Faster R-CNN based fish head detector. The system reaches a MOTA of up to 77.6%. Links to the code and dataset is available at the project page http://vap.aau.dk/3d-zef
****

X3D: Expanding Architectures for Efficient Video Recognition
Christoph Feichtenhofer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 203-213
This paper presents X3D, a family of efficient video networks that progressively expand a tiny 2D image classification architecture along multiple network axes, in space, time, width and depth. Inspired by feature selection methods in machine learning, a simple stepwise network expansion approach is employed that expands a single axis in each step, such that good accuracy to complexity trade-off i

s achieved. To expand X3D to a specific target complexity, we perform progressive forward expansion followed by backward contraction. X3D achieves state-of-the-art performance while requiring 4.8x and 5.5x fewer multiply-adds and parameters for similar accuracy as previous work. Our most surprising finding is that networks with high spatiotemporal resolution can perform well, while being extremely light in terms of network width and parameters. We report competitive accuracy at unprecedented efficiency on video classification and detection benchmarks. Code is available at: https://github.com/facebookresearch/SlowFast.

****

Unsupervised Intra-Domain Adaptation for Semantic Segmentation Through Self-Supervision

Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3764-3773

Convolutional neural network-based approaches have achieved remarkable progress in semantic segmentation. However, these approaches heavily rely on annotated data which are labor intensive. To cope with this limitation, automatically annotated data generated from graphic engines are used to train segmentation models. However, the models trained from synthetic data are difficult to transfer to real images. To tackle this issue, previous works have considered directly adapting models from the source data to the unlabeled target data (to reduce the inter-domain gap). Nonetheless, these techniques do not consider the large distribution gap among the target data itself (intra-domain gap). In this work, we propose a two-step self-supervised domain adaptation approach to minimize the inter-domain and intra-domain gap together. First, we conduct the inter-domain adaptation of the model, from this adaptation, we separate target domain into an easy and hard split using an entropy-based ranking function. Finally, to decrease the intra-domain gap, we propose to employ a self-supervised adaptation technique from the easy to the hard subdomain. Experimental results on numerous benchmark datasets highlight the effectiveness of our method against existing state-of-the-art approaches. The source code is available at https://github.com/feipan664/IntraDA.git.

****

Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks

Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, Trevor Darrell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1049-1059

Human action is naturally compositional: humans can easily recognize and perform actions with objects that are different from those used in training demonstrations. In this paper, we study the compositionality of action by looking into the dynamics of subject-object interactions. We propose a novel model which can explicitly reason about the geometric relations between constituent objects and an agent performing an action. To train our model, we collect dense object box annotations on the Something-Something dataset. We propose a novel compositional action recognition task where the training combinations of verbs and nouns do not overlap with the test set. The novel aspects of our model are applicable to activities with prominent object interaction dynamics and to objects which can be tracked using state-of-the-art approaches; for activities without clearly defined spatial object-agent interactions, we rely on baseline scene-level spatio-temporal representations. We show the effectiveness of our approach not only on the proposed compositional action recognition task but also in a few-shot compositional setting which requires the model to generalize across both object appearance and action category.

****

Exploring Spatial-Temporal Multi-Frequency Analysis for High-Fidelity and Temporal-Consistency Video Prediction

Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, Xiaowei Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4554-4563

Video prediction is a pixel-wise dense prediction task to infer future frames based on past frames. Missing appearance details and motion blur are still two major problems for current models, leading to image distortion and temporal inconsistency. We point out the necessity of exploring multi-frequency analysis to deal with the two problems. Inspired by the frequency band decomposition characteristic of Human Vision System (HVS), we propose a video prediction network based on multi-level wavelet analysis to uniformly deal with spatial and temporal information. Specifically, multi-level spatial discrete wavelet transform decomposes each video frame into anisotropic sub-bands with multiple frequencies, helping to enrich structural information and reserve fine details. On the other hand, multilevel temporal discrete wavelet transform which operates on time axis decomposes the frame sequence into sub-band groups of different frequencies to accurately capture multifrequency motions under a fixed frame rate. Extensive experiments on diverse datasets demonstrate that our model shows significant improvements on fidelity and temporal consistency over the state-of-the-art works. Source code and videos are available at https://github.com/Bei-Jin/STMFANet.
****

Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud
Weijing Shi, Raj Rajkumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1711-1719
In this paper, we propose a graph neural network to detect objects from a LiDAR point cloud. Towards this end, we encode the point cloud efficiently in a fixed radius near-neighbors graph. We design a graph neural network, named Point-GNN, to predict the category and shape of the object that each vertex in the graph belongs to. In Point-GNN, we propose an auto-registration mechanism to reduce translation variance, and also design a box merging and scoring operation to combine detections from multiple vertices accurately. Our experiments on the KITTI benchmark show the proposed approach achieves leading accuracy using the point cloud alone and can even surpass fusion-based algorithms. Our results demonstrate the potential of using the graph neural network as a new approach for 3D object detection. The code is available at https://github.com/WeijingShi/Point-GNN.
****

How Much Time Do You Have? Modeling Multi-Duration Saliency
Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, Zoya Bylinskii; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4473-4482
What jumps out in a single glance of an image is different than what you might notice after closer inspection. Yet conventional models of visual saliency produce predictions at an arbitrary, fixed viewing duration, offering a limited view of the rich interactions between image content and gaze location. In this paper we propose to capture gaze as a series of snapshots, by generating population-level saliency heatmaps for multiple viewing durations. We collect the CodeCharts1K dataset, which contains multiple distinct heatmaps per image corresponding to 0.5, 3, and 5 seconds of free-viewing. We develop an LSTM-based model of saliency that simultaneously trains on data from multiple viewing durations. Our Multi-Duration Saliency Excited Model (MD-SEM) achieves competitive performance on the LSUN 2017 Challenge with 57% fewer parameters than comparable architectures. It is the first model that produces heatmaps at multiple viewing durations, enabling applications where multi-duration saliency can be used to prioritize visual content to keep, transmit, and render.
****

Learning Fused Pixel and Feature-Based View Reconstructions for Light Fields
Jinglei Shi, Xiaoran Jiang, Christine Guillemot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2555-2564
In this paper, we present a learning-based framework for light field view synthesis from a subset of input views. Building upon a light-weight optical flow estimation network to obtain depth maps, our method employs two reconstruction modules in pixel and feature domains respectively. For the pixel-wise reconstruction, occlusions are explicitly handled by a disparity-dependent interpolation filter, whereas inpainting on disoccluded areas is learned by convolutional layers. Du

e to disparity inconsistencies, the pixel-based reconstruction may lead to blurr iness in highly textured areas as well as on object contours. On the contrary, t he feature-based reconstruction well performs on high frequencies, making the re construction in the two domains complementary. End-to-end learning is finally pe rformed including a fusion module merging pixel and feature-based reconstruction s. Experimental results show that our method achieves state-of-the-art performan ce on both synthetic and real-world datasets, moreover, it is even able to exten d light fields' baseline by extrapolating high quality views without additional training.
****

Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recogni tion and Retrieval
Tobias Weyand, Andre Araujo, Bingyi Cao, Jack Sim; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2575-2 584
While image retrieval and instance recognition techniques are progressing rapidl y, there is a need for challenging datasets to accurately measure their performa nce -- while posing novel challenges that are relevant for practical application s. We introduce the Google Landmarks Dataset v2 (GLDv2), a new benchmark for lar ge-scale, fine-grained instance recognition and image retrieval in the domain of human-made and natural landmarks. GLDv2 is the largest such dataset to date by a large margin, including over 5M images and 200k distinct instance labels. Its test set consists of 118k images with ground truth annotations for both the retr ieval and recognition tasks. The ground truth construction involved over 800 hou rs of human annotator work. Our new dataset has several challenging properties i nspired by real-world applications that previous datasets did not consider: An e xtremely long-tailed class distribution, a large fraction of out-of-domain test photos and large intra-class variability. The dataset is sourced from Wikimedia Commons, the world's largest crowdsourced collection of landmark photos. We prov ide baseline results for both recognition and retrieval tasks based on state-of- the-art methods as well as competitive results from a public challenge. We furth er demonstrate the suitability of the dataset for transfer learning by showing t hat image embeddings trained on it achieve competitive retrieval performance on independent datasets. The dataset images, ground-truth and metric scoring code a re available at https://github.com/cvdfoundation/google-landmark
****

Lightweight Photometric Stereo for Facial Details Recovery
Xueying Wang, Yudong Guo, Bailin Deng, Juyong Zhang; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 740- 749
Recently, 3D face reconstruction from a single image has achieved great success with the help of deep learning and shape prior knowledge, but they often fail to produce accurate geometry details. On the other hand, photometric stereo method s can recover reliable geometry details, but require dense inputs and need to so lve a complex optimization problem. In this paper, we present a lightweight stra tegy that only requires sparse inputs or even a single image to recover high-fid elity face shapes with images captured under near-field lights. To this end, we construct a dataset containing 84 different subjects with 29 expressions under 3 different lights. Data augmentation is applied to enrich the data in terms of d iversity in identity, lighting, expression, etc. With this constructed dataset, we propose a novel neural network specially designed for photometric stereo base d 3D face reconstruction. Extensive experiments and comparisons demonstrate that our method can generate high-quality reconstruction results with one to three f acial images captured under near-field lights. Our full framework is available a t https://github.com/Juyong/FacePSNet.
****

SAL: Sign Agnostic Learning of Shapes From Raw Data
Matan Atzmon, Yaron Lipman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2565-2574
Recently, neural networks have been used as implicit representations for surface

reconstruction, modelling, learning, and generation. So far, training neural networks to be implicit representations of surfaces required training data sampled from a ground-truth signed implicit functions such as signed distance or occupancy functions, which are notoriously hard to compute. In this paper we introduce Sign Agnostic Learning (SAL), a deep learning approach for learning implicit shape representations directly from raw, unsigned geometric data, such as point clouds and triangle soups. We have tested SAL on the challenging problem of surface reconstruction from an un-oriented point cloud, as well as end-to-end human shape space learning directly from raw scans dataset, and achieved state of the art reconstructions compared to current approaches. We believe SAL opens the door to many geometric deep learning applications with real-world data, alleviating the usual painstaking, often manual pre-process.
****

Learning Filter Pruning Criteria for Deep Convolutional Neural Networks Acceleration
Yang He, Yuhang Ding, Ping Liu, Linchao Zhu, Hanwang Zhang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2009-2018
Filter pruning has been widely applied to neural network compression and acceleration. Existing methods usually utilize pre-defined pruning criteria, such as Lp-norm, to prune unimportant filters. There are two major limitations to these methods. First, existing methods fail to consider the variety of filter distribution across layers. To extract features of the coarse level to the fine level, the filters of different layers have various distributions. Therefore, it is not suitable to utilize the same pruning criteria to different functional layers. Second, prevailing layer-by-layer pruning methods process each layer independently and sequentially, failing to consider that all the layers in the network collaboratively make the final prediction. In this paper, we propose Learning Filter Pruning Criteria (LFPC) to solve the above problems. Specifically, we develop a differentiable pruning criteria sampler. This sampler is learnable and optimized by the validation loss of the pruned network obtained from the sampled criteria. In this way, we could adaptively select the appropriate pruning criteria for different functional layers. Besides, when evaluating the sampled criteria, LFPC comprehensively consider the contribution of all the layers at the same time. Experiments validate our approach on three image classification benchmarks. Notably, on ILSVRC-2012, our LFPC reduces more than 60% FLOPs on ResNet-50 with only 0.83% top-5 accuracy loss.
****

WCP: Worst-Case Perturbations for Semi-Supervised Deep Learning
Liheng Zhang, Guo-Jun Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3912-3921
In this paper, we present a novel regularization mechanism for training deep networks by minimizing the Worse-Case Perturbation (WCP). It is based on the idea that a robust model is least likely to be affected by small perturbations, such that its output decisions should be as stable as possible on both labeled and unlabeled examples. We will consider two forms of WCP regularizations -- additive and DropConnect perturbations, which impose additive noises on network weights, and make structural changes by dropping the network connections, respectively. We will show that the worse cases of both perturbations can be derived by solving respective optimization problems with spectral methods. The WCP can be minimized on both labeled and unlabeled data so that networks can be trained in a semi-supervised fashion. This leads to a novel paradigm of semi-supervised classifiers by stabilizing the predicted outputs in presence of the worse-case perturbations imposed on the network weights and structures.
****

Separating Particulate Matter From a Single Microscopic Image
Tushar Sandhan, Jin Young Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4584-4593
Particulate matter (PM) is the blend of various solid and liquid particles suspended in atmosphere. These submicron particles are imperceptible for usual hand-h

eld camera photography, but become a great obstacle in microscopic imaging. PM r
emoval from a single microscopic image is a highly ill-posed and one of the chal
lenging image denoising problems. In this work, we thoroughly analyze the physic
al properties of PM, microscope and their inevitable interaction; and propose an
 optimization scheme, which removes the PM from a high-resolution microscopic im
age within a few seconds. Experiments on real world microscopic images show that
 the proposed method significantly outperforms other competitive image denoising
 methods. It preserves the comprehensive microscopic foreground details while cl
early separating the PM from a single monochromatic or color image.
****

Event Probability Mask (EPM) and Event Denoising Convolutional Neural Network (E
DnCNN) for Neuromorphic Cameras
R. Wes Baldwin, Mohammed Almatrafi, Vijayan Asari, Keigo Hirakawa; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2020, pp. 1701-1710
This paper presents a novel method for labeling real-world neuromorphic camera s
ensor data by calculating the likelihood of generating an event at each pixel wi
thin a short time window, which we refer to as "event probability mask" or EPM.
Its applications include (i) objective benchmarking of event denoising performan
ce, (ii) training convolutional neural networks for noise removal called "event
denoising convolutional neural network" (EDnCNN), and (iii) estimating internal
neuromorphic camera parameters. We provide the first dataset (DVSNOISE20) of rea
l-world labeled neuromorphic camera events for noise removal.
****

Peek-a-Boo: Occlusion Reasoning in Indoor Scenes With Plane Representations
Ziyu Jiang, Buyu Liu, Samuel Schulter, Zhangyang Wang, Manmohan Chandraker;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2020, pp. 113-121
We address the challenging task of occlusion-aware indoor 3D scene understanding
. We represent scenes by a set of planes, where each one is defined by its norma
l, offset and two masks outlining (i) the extent of the visible part and (ii) th
e full region that consists of both visible and occluded parts of the plane. We
infer these planes from a single input image with a novel neural network archite
cture. It consists of a two-branch category-specific module that aims to predict
 layout and objects of the scene separately so that different types of planes ca
n be handled better. We also introduce a novel loss function based on plane warp
ing that can leverage multiple views at training time for improved occlusion-awa
re reasoning. In order to train and evaluate our occlusion-reasoning model, we u
se the ScanNet dataset and propose (i) a strategy to automatically extract groun
d truth for both visible and hidden regions and (ii) a new evaluation metric tha
t specifically focuses on the prediction in hidden regions. We empirically demon
strate that our proposed approach can achieve higher accuracy for occlusion reas
oning compared to competitive baselines on the ScanNet dataset, e.g. 42.65% rela
tive improvement on hidden regions.
****

Unsupervised Learning of Probably Symmetric Deformable 3D Objects From Images in
 the Wild
Shangzhe Wu, Christian Rupprecht, Andrea Vedaldi; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1-10
We propose a method to learn 3D deformable object categories from raw single-vie
w images, without external supervision. The method is based on an autoencoder th
at factors each input image into depth, albedo, viewpoint and illumination. In o
rder to disentangle these components without supervision, we use the fact that m
any object categories have, at least in principle, a symmetric structure. We sho
w that reasoning about illumination allows us to exploit the underlying object s
ymmetry even if the appearance is not symmetric due to shading. Furthermore, we
model objects that are probably, but not certainly, symmetric by predicting a sy
mmetry probability map, learned end-to-end with the other components of the mode
l. Our experiments show that this method can recover very accurately the 3D shap
e of human faces, cat faces and cars from single-view images, without any superv

ision or a prior shape model. On benchmarks, we demonstrate superior accuracy co mpared to another method that uses supervision at the level of 2D image correspo ndences.
****

Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classifi cation with Unannotated Histopathological Images
Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, Ichiro Takeu chi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2020, pp. 3852-3861
We propose a new method for cancer subtype classification from histopathological images, which can automatically detect tumor-specific features in a given whole slide image (WSI). The cancer subtype should be classified by referring to a WS I, i.e., a large-sized image (typically 40,000x40,000 pixels) of an entire patho logical tissue slide, which consists of cancer and non-cancer portions. One diff iculty arises from the high cost associated with annotating tumor regions in WSI s. Furthermore, both global and local image features must be extracted from the WSI by changing the magnifications of the image. In addition, the image features should be stably detected against the differences of staining conditions among the hospitals/specimens. In this paper, we develop a new CNN-based cancer subtyp e classification method by effectively combining multiple-instance, domain adver sarial, and multi-scale learning frameworks in order to overcome these practical difficulties. When the proposed method was applied to malignant lymphoma subtyp e classifications of 196 cases collected from multiple hospitals, the classifica tion performance was significantly better than the standard CNN or other convent ional methods, and the accuracy compared favorably with that of standard patholo gists.
****

STAViS: Spatio-Temporal AudioVisual Saliency Network
Antigoni Tsiami, Petros Koutras, Petros Maragos; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4766-4776
We introduce STAViS, a spatio-temporal audiovisual saliency network that combine s spatio-temporal visual and auditory information in order to efficiently addres s the problem of saliency estimation in videos. Our approach employs a single ne twork that combines visual saliency and auditory features and learns to appropri ately localize sound sources and to fuse the two saliencies in order to obtain a final saliency map. The network has been designed, trained end-to-end, and eval uated on six different databases that contain audiovisual eye-tracking data of a large variety of videos. We compare our method against 8 different state-of-the -art visual saliency models. Evaluation results across databases indicate that o ur STAViS model outperforms our visual only variant as well as the other state-o f-the-art models in the majority of cases. Also, the consistently good performan ce it achieves for all databases indicates that it is appropriate for estimating saliency "in-the-wild". The code is available at https://github.com/atsiami/STA ViS.
****

OctSqueeze: Octree-Structured Entropy Model for LiDAR Compression
Lila Huang, Shenlong Wang, Kelvin Wong, Jerry Liu, Raquel Urtasun; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1313-1323
We present a novel deep compression algorithm to reduce the memory footprint of LiDAR point clouds. Our method exploits the sparsity and structural redundancy b etween points to reduce the bitrate. Towards this goal, we first encode the poin t cloud into an octree, a data-efficient structure suitable for sparse point clo uds. We then design a tree-structured conditional entropy model that can be dire ctly applied to octree structures to predict the probability of a symbol's occur rence. We validate the effectiveness of our method over two large-scale datasets . The results demonstrate that our approach reduces the bitrate by 10- 20% at th e same reconstruction quality, compared to the previous state-of-the-art. Import antly, we also show that for the same bitrate, our approach outperforms other co

mpression algorithms when performing downstream 3D segmentation and detection tasks using compressed representations. This helps advance the feasibility of using point cloud compression to reduce the onboard and offboard storage for safety-critical applications such as self-driving cars, where a single vehicle captures 84 billion points per day.
****

Block-Wisely Supervised Neural Architecture Search With Knowledge Distillation
Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, Xiaojun Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1989-1998
Neural Architecture Search (NAS), aiming at automatically designing network architectures by machines, is expected to bring about a new revolution in machine learning. Despite these high expectation, the effectiveness and efficiency of existing NAS solutions are unclear, with some recent works going so far as to suggest that many existing NAS solutions are no better than random architecture selection. The ineffectiveness of NAS solutions may be attributed to inaccurate architecture evaluation. Specifically, to speed up NAS, recent works have proposed under-training different candidate architectures in a large search space concurrently by using shared network parameters; however, this has resulted in incorrect architecture ratings and furthered the ineffectiveness of NAS. In this work, we propose to modularize the large search space of NAS into blocks to ensure that the potential candidate architectures are fully trained; this reduces the representation shift caused by the shared parameters and leads to the correct rating of the candidates. Thanks to the block-wise search, we can also evaluate all of the candidate architectures within each block. Moreover, we find that the knowledge of a network model lies not only in the network parameters but also in the network architecture. Therefore, we propose to distill the neural architecture (DNA) knowledge from a teacher model to supervise our block-wise architecture search, which significantly improves the effectiveness of NAS. Remarkably, the performance of our searched architectures has exceeded the teacher model, demonstrating the practicability of our method. Finally, our method achieves a state-of-the-art 78.4% top-1 accuracy on ImageNet in a mobile setting. All of our searched models along with the evaluation code are available at https://github.com/changlin31/DNA.
****

PANDA: A Gigapixel-Level Human-Centric Video Dataset
Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3268-3278
We present PANDA, the first gigaPixel-level humAN-centric viDeo dAtaset, for large-scale, long-term, and multi-object visual analysis. The videos in PANDA were captured by a gigapixel camera and cover real-world scenes with both wide field-of-view (1 square kilometer area) and high-resolution details (gigapixel-level /frame). The scenes may contain 4k head counts with over 100x scale variation. PANDA provides enriched and hierarchical ground-truth annotations, including 15,974.6k bounding boxes, 111.8k fine-grained attribute labels, 12.7k trajectories, 2.2k groups and 2.9k interactions. We benchmark the human detection and tracking tasks. Due to the vast variance of pedestrian pose, scale, occlusion and trajectory, existing approaches are challenged by both accuracy and efficiency. Given the uniqueness of PANDA with both wide FoV and high resolution, a new task of interaction-aware group detection is introduced. We design a 'global-to-local zoom-in' framework, where global trajectories and local interactions are simultaneously encoded, yielding promising results. We believe PANDA will contribute to the community of artificial intelligence and praxeology by understanding human behaviors and interactions in large-scale real-world scenes. PANDA Website: http://www.panda-dataset.com.
****

Few-Shot Learning of Part-Specific Probability Space for 3D Shape Segmentation
Lingjing Wang, Xiang Li, Yi Fang; Proceedings of the IEEE/CVF Conference on Co

mputer Vision and Pattern Recognition (CVPR), 2020, pp. 4504-4513
Recently, deep neural networks are introduced as supervised discriminative models for the learning of 3D point cloud segmentation. Most previous supervised methods require a large number of training data with human annotation part labels to guide the training process to ensure the model's generalization abilities on test data. In comparison, we propose a novel 3D shape segmentation method that requires few labeled data for training. Given an input 3D shape, the training of our model starts with identifying a similar 3D shape with part annotations from a mini-pool of shape templates (e.g. 10 shapes). With the selected template shape, a novel Coherent Point Transformer is proposed to fully leverage the power of a deep neural network to smoothly morph the template shape towards the input shape. Then, based on the transformed template shapes with part labels, a newly proposed Part-specific Density Estimator is developed to learn a continuous part-specific probability distribution function on the entire 3D space with a batch consistency regularization term. With the learned part-specific probability distribution, our model is able to predict the part labels of a new input 3D shape in an end-to-end manner. We demonstrate that our proposed method can achieve remarkable segmentation results on the ShapeNet dataset with few shots, compared to previous supervised learning approaches.
****

Actor-Transformers for Group Activity Recognition
Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 839-848
This paper strives to recognize individual actions and group activities from videos. While existing solutions for this challenging problem explicitly model spatial and temporal relationships based on location of individual actors, we propose an actor-transformer model able to learn and selectively extract information relevant for group activity recognition. We feed the transformer with rich actor-specific static and dynamic representations expressed by features from a 2D pose network and 3D CNN, respectively. We empirically study different ways to combine these representations and show their complementary benefits. Experiments show what is important to transform and how it should be transformed. What is more, actor-transformers achieve state-of-the-art results on two publicly available benchmarks for group activity recognition, outperforming the previous best published results by a considerable margin
****

Domain Adaptation for Image Dehazing
Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2808-2817
Image dehazing using learning-based methods has achieved state-of-the-art performance in recent years. However, most existing methods train a dehazing model on synthetic hazy images, which are less able to generalize well to real hazy images due to domain shift. To address this issue, we propose a domain adaptation paradigm, which consists of an image translation module and two image dehazing modules. Specifically, we first apply a bidirectional translation network to bridge the gap between the synthetic and real domains by translating images from one domain to another. And then, we use images before and after translation to train the proposed two image dehazing networks with a consistency constraint. In this phase, we incorporate the real hazy image into the dehazing training via exploiting the properties of the clear image (e.g., dark channel prior and image gradient smoothing) to further improve the domain adaptivity. By training image translation and dehazing network in an end-to-end manner, we can obtain better effects of both image translation and dehazing. Experimental results on both synthetic and real-world images demonstrate that our model performs favorably against the state-of-the-art dehazing algorithms.
****

Exploiting Joint Robustness to Adversarial Perturbations
Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Jeremy Dawson, Nasser M.

Nasrabadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1122-1131

Recently, ensemble models have demonstrated empirical capabilities to alleviate the adversarial vulnerability. In this paper, we exploit first-order interactions within ensembles to formalize a reliable and practical defense. We introduce a scenario of interactions that certifiably improves the robustness according to the size of the ensemble, the diversity of the gradient directions, and the balance of the member's contribution to the robustness. We present a joint gradient phase and magnitude regularization (GPMR) as a vigorous approach to impose the desired scenario of interactions among members of the ensemble. Through extensive experiments, including gradient-based and gradient-free evaluations on several datasets and network architectures, we validate the practical effectiveness of the proposed approach compared to the previous methods. Furthermore, we demonstrate that GPMR is orthogonal to other defense strategies developed for single classifiers and their combination can further improve the robustness of ensembles.

****