Learning Video Representations From Correspondence Proposals
Xingyu Liu, Joon-Young Lee, Hailin Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4273-4281
Correspondences between frames encode rich information about dynamic content in videos. However, it is challenging to effectively capture and learn those due to their irregular structure and complex dynamics. In this paper, we propose a novel neural network that learns video representations by aggregating information from potential correspondences. This network, named CPNet, can learn evolving 2D fields with temporal consistency. In particular, it can effectively learn representations for videos by mixing appearance and long-range motion with an RGB-only input. We provide extensive ablation experiments to validate our model. CPNet shows stronger performance than existing methods on Kinetics and achieves the state-of-the-art performance on Something-Something and Jester. We provide analysis towards the behavior of our model and show its robustness to errors in proposals.
****
SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks
Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4282-4291
Siamese network based trackers formulate tracking as convolutional feature cross-correlation between target template and searching region. However, Siamese trackers still have accuracy gap compared with state-of-the-art algorithms and they cannot take advantage of feature from deep networks, such as ResNet-50 or deeper. In this work we prove the core reason comes from the lack of strict translation invariance. By comprehensive theoretical analysis and experimental validations, we break this restriction through a simple yet effective spatial aware sampling strategy and successfully train a ResNet-driven Siamese tracker with significant performance gain. Moreover, we propose a new model architecture to perform depth-wise and layer-wise aggregations, which not only further improves the accuracy but also reduces the model size. We conduct extensive ablation studies to demonstrate the effectiveness of the proposed tracker, which obtains currently the best results on four large tracking benchmarks, including OTB2015, VOT2018, UAV123, and LaSOT. Our model will be released to facilitate further studies based on this problem.
****
Sphere Generative Adversarial Network Based on Geometric Moment Matching
Sung Woo Park, Junseok Kwon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4292-4301
We propose sphere generative adversarial network (GAN), a novel integral probability metric (IPM)-based GAN. Sphere GAN uses the hypersphere to bound IPMs in the objective function. Thus, it can be trained stably. On the hypersphere, sphere GAN exploits the information of higher-order statistics of data using geometric moment matching, thereby providing more accurate results. In the paper, we mathematically prove the good properties of sphere GAN. In experiments, sphere GAN quantitatively and qualitatively surpasses recent state-of-the-art GANs for unsupervised image generation problems with the CIFAR-10, STL-10, and LSUN bedroom datasets. Source code is available at https://github.com/pswkiki/SphereGAN.
****
Adversarial Attacks Beyond the Image Space
Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4302-4311
Generating adversarial examples is an intriguing problem and an important way of understanding the working mechanism of deep neural networks. Most existing approaches generated perturbations in the image space, i.e., each pixel can be modified independently. However, in this paper we pay special attention to the subset of adversarial examples that correspond to meaningful changes in 3D physical properties (like rotation and translation, illumination condition, etc.). These adversaries arguably pose a more serious concern, as they demonstrate the possibil

ity of causing neural network failure by easy perturbations of real-world 3D obj ects and scenes. In the contexts of object classification and visual question a nswering, we augment state-of-the-art deep neural networks that receive 2D input images with a rendering module (either differentiable or not) in front, so that a 3D scene (in the physical space) is rendered into a 2D image (in the image sp ace), and then mapped to a prediction (in the output space). The adversarial per turbations can now go beyond the image space, and have clear meanings in the 3D physical world. Though image-space adversaries can be interpreted as per-pixel a lbedo change, we verify that they cannot be well explained along these physicall y meaningful dimensions, which often have a non-local effect. But it is still po ssible to successfully attack beyond the image space on the physical space, thou gh this is more difficult than image-space attacks, reflected in lower success r ates and heavier perturbations required.

****

Evading Defenses to Transferable Adversarial Examples by Translation-Invariant A ttacks

Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4312-4321

Deep neural networks are vulnerable to adversarial examples, which can mislead c lassifiers by adding imperceptible perturbations. An intriguing property of adve rsarial examples is their good transferability, making black-box attacks feasibl e in real-world applications. Due to the threat of adversarial attacks, many met hods have been proposed to improve the robustness. Several state-of-the-art defe nses are shown to be robust against transferable adversarial examples. In this p aper, we propose a translation-invariant attack method to generate more transfer able adversarial examples against the defense models. By optimizing a perturbati on over an ensemble of translated images, the generated adversarial example is l ess sensitive to the white-box model being attacked and has better transferabili ty. To improve the efficiency of attacks, we further show that our method can be implemented by convolving the gradient at the untranslated image with a pre-def ined kernel. Our method is generally applicable to any gradient-based attack met hod. Extensive experiments on the ImageNet dataset validate the effectiveness of the proposed method. Our best attack fools eight state-of-the-art defenses at a n 82% success rate on average based only on the transferability, demonstrating t he insecurity of the current defense techniques.

****

Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attack s and Defenses

Jerome Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sab ourin, Eric Granger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4322-4330

Research on adversarial examples in computer vision tasks has shown that small, often imperceptible changes to an image can induce misclassification, which has security implications for a wide range of image processing systems. Considering L2 norm distortions, the Carlini and Wagner attack is presently the most effecti ve white-box attack in the literature. However, this method is slow since it per forms a line-search for one of the optimization terms, and often requires thousa nds of iterations. In this paper, an efficient approach is proposed to generate gradient-based attacks that induce misclassifications with low L2 norm, by decou pling the direction and the norm of the adversarial perturbation that is added t o the image. Experiments conducted on the MNIST, CIFAR-10 and ImageNet datasets indicate that our attack achieves comparable results to the state-of-the-art (in terms of L2 norm) with considerably fewer iterations (as few as 100 iterations) , which opens the possibility of using these attacks for adversarial training. M odels trained with our attack achieve state-of-the-art robustness against white- box gradient-based L2 attacks on the MNIST and CIFAR-10 datasets, outperforming the Madry defense when the attacks are limited to a maximum norm.

****

A General and Adaptive Robust Loss Function

Jonathan T. Barron; Proceedings of the IEEE/CVF Conference on Computer Vision an

d Pattern Recognition (CVPR), 2019, pp. 4331-4339
We present a generalization of the Cauchy/Lorentzian, Geman-McClure, Welsch/Lecl erc, generalized Charbonnier, Charbonnier/pseudo-Huber/L1-L2, and L2 loss functi ons. By introducing robustness as a continuous parameter, our loss function allo ws algorithms built around robust loss minimization to be generalized, which imp roves performance on basic vision tasks such as registration and clustering. Int erpreting our loss as the negative log of a univariate density yields a general probability distribution that includes normal and Cauchy distributions as specia l cases. This probabilistic interpretation enables the training of neural networ ks in which the robustness of the loss automatically adapts itself during traini ng, which improves performance on learning-based tasks such as generative image synthesis and unsupervised monocular depth estimation, without requiring any man ual parameter tuning.
****

Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Accel eration
Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, Yi Yang; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4340-4349
Previous works utilized "smaller-norm-less-important" criterion to prune filters with smaller norm values in a convolutional neural network. In this paper, we analyze this norm-based criterion and point out that its effectiveness depends o n two requirements that are not always met: (1) the norm deviation of the filte rs should be large; (2) the minimum norm of the filters should be small. To solv e this problem, we propose a novel filter pruning method, namely Filter Pruning via Geometric Median (FPGM), to compress the model regardless of those two requi rements. Unlike previous methods, FPGM compresses CNN models by pruning filters with redundancy, rather than those with"relatively less" importance. When applie d to two image classification benchmarks, our method validates its usefulness an d strengths. Notably, on CIFAR-10, FPGM reduces more than 52% FLOPs on ResNet-11 0 with even 2.69% relative accuracy improvement. Moreover, on ILSVRC-2012, FPGM reduces more than 42% FLOPs on ResNet-101 without top-5 accuracy drop, which has advanced the state-of-the-art. Code is publicly available on GitHub: https://gi thub.com/he-y/filter-pruning-geometric-median
****

Learning to Quantize Deep Networks by Optimizing Quantization Intervals With Tas k Loss
Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngju n Kwak, Sung Ju Hwang, Changkyu Choi; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4350-4359
Reducing bit-widths of activations and weights of deep networks makes it efficie nt to compute and store them in memory, which is crucial in their deployments to resource-limited devices, such as mobile phones. However, decreasing bit-widths with quantization generally yields drastically degraded accuracy. To tackle thi s problem, we propose to learn to quantize activations and weights via a trainab le quantizer that transforms and discretizes them. Specifically, we parameterize the quantization intervals and obtain their optimal values by directly minimizi ng the task loss of the network. This quantization-interval-learning (QIL) allow s the quantized networks to maintain the accuracy of the full-precision (32-bit) networks with bit-width as low as 4-bit and minimize the accuracy degeneration with further bit-width reduction (i.e., 3 and 2-bit). Moreover, our quantizer ca n be trained on a heterogeneous dataset, and thus can be used to quantize pretra ined networks without access to their training data. We demonstrate the effectiv eness of our trainable quantizer on ImageNet dataset with various network archit ectures such as ResNet-18, -34 and AlexNet, on which it outperforms existing met hods to achieve the state-of-the-art accuracy.
****

Not All Areas Are Equal: Transfer Learning for Semantic Segmentation via Hierarc hical Region Selection
Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, Lizhuang Ma; P

roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4360-4369
The success of deep neural networks for semantic segmentation heavily relies on large-scale and well-labeled datasets, which are hard to collect in practice. Synthetic data offers an alternative to obtain ground-truth labels for free. However, models directly trained on synthetic data often struggle to generalize to real images. In this paper, we consider transfer learning for semantic segmentation that aims to mitigate the gap between abundant synthetic data (source domain) and limited real data (target domain). Unlike previous approaches that either learn mappings to target domain or finetune on target images, our proposed method jointly learn from real images and selectively from realistic pixels in synthetic images to adapt to the target domain. Our key idea is to have weighting networks to score how similar the synthetic pixels are to real ones, and learn such weighting at pixel-, region- and image-levels. We jointly learn these hierarchical weighting networks and segmentation network in an end-to-end manner. Extensive experiments demonstrate that our proposed approach significantly outperforms other existing baselines, and is applicable to scenarios with extremely limited real images.
****

Unsupervised Learning of Dense Shape Correspondence
Oshri Halimi, Or Litany, Emanuele Rodola, Alex M. Bronstein, Ron Kimmel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4370-4379
We introduce the first completely unsupervised correspondence learning approach for deformable 3D shapes. Key to our model is the understanding that natural deformations (such as changes in pose) approximately preserve the metric structure of the surface, yielding a natural criterion to drive the learning process toward distortion-minimizing predictions. On this basis, we overcome the need for annotated data and replace it by a purely geometric criterion. The resulting learning model is class-agnostic, and is able to leverage any type of deformable geometric data for the training phase. In contrast to existing supervised approaches which specialize on the class seen at training time, we demonstrate stronger generalization as well as applicability to a variety of challenging settings. We showcase our method on a wide selection of correspondence benchmarks, where we outperform other methods in terms of accuracy, generalization, and efficiency.
****

Unsupervised Visual Domain Adaptation: A Deep Max-Margin Gaussian Process Approach
Minyoung Kim, Pritish Sahu, Behnam Gholami, Vladimir Pavlovic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4380-4390
For unsupervised domain adaptation, the target domain error can be provably reduced by having a shared input representation that makes the source and target domains indistinguishable from each other. Very recently it has been shown that it is not only critical to match the marginal input distributions, but also align the output class distributions. The latter can be achieved by minimizing the maximum discrepancy of predictors. In this paper, we take this principle further by proposing a more systematic and effective way to achieve hypothesis consistency using Gaussian processes (GP). The GP allows us to induce a hypothesis space of classifiers from the posterior distribution of the latent random functions, turning the learning into a large-margin posterior separation problem, significantly easier to solve than previous approaches based on adversarial minimax optimization. We formulate a learning objective that effectively influences the posterior to minimize the maximum discrepancy. This is shown to be equivalent to maximizing margins and minimizing uncertainty of the class predictions in the target domain. Empirical results demonstrate that our approach leads to state-to-the-art performance superior to existing methods on several challenging benchmarks for domain adaptation.
****

Balanced Self-Paced Learning for Generative Adversarial Clustering Network

Kamran Ghasedi, Xiaoqian Wang, Cheng Deng, Heng Huang; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 43 91-4400

Clustering is an important problem in various machine learning applications, but still a challenging task when dealing with complex real data. The existing clus tering algorithms utilize either shallow models with insufficient capacity for c apturing the non-linear nature of data, or deep models with large number of para meters prone to overfitting. In this paper, we propose a deep Generative Advers arial Clustering Network (ClusterGAN), which tackles the problems of training of deep clustering models in unsupervised manner. ClusterGAN consists of three net works, a discriminator, a generator and a clusterer (i.e. a clustering network). We employ an adversarial game between these three players to synthesize realist ic samples given discriminative latent variables via the generator, and learn th e inverse mapping of the real samples to the discriminative embedding space via the clusterer. Moreover, we utilize a conditional entropy minimization loss to i ncrease/decrease the similarity of intra/inter cluster samples. Since the ground -truth similarities are unknown in clustering task, we propose a novel balanced self-paced learning algorithm to gradually include samples into training from ea sy to difficult, while considering the diversity of selected samples from all cl usters. Therefore, our method makes it possible to efficiently train clusterers with large depth by leveraging the proposed adversarial game and balanced self-p aced learning algorithm. According our experiments, ClusterGAN achieves competit ive results compared to the state-of-the-art clustering and hashing models on se veral datasets.

****

A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras, Samuli Laine, Timo Aila; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401-4410

We propose an alternative generator architecture for generative adversarial netw orks, borrowing from style transfer literature. The new architecture leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the g enerated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis. The new generator improves the state-of-the-art in te rms of traditional distribution quality metrics, leads to demonstrably better in terpolation properties, and also better disentangles the latent factors of varia tion. To quantify interpolation quality and disentanglement, we propose two new, automated methods that are applicable to any generator architecture. Finally, w e introduce a new, highly varied and high-quality dataset of human faces.

****

Parallel Optimal Transport GAN

Gil Avraham, Yan Zuo, Tom Drummond; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4411-4420

Although Generative Adversarial Networks (GANs) are known for their sharp realis m in image generation, they often fail to estimate areas of the data density. Th is leads to low modal diversity and at times distorted generated samples. These problems essentially arise from poor estimation of the distance metric responsib le for training these networks. To address these issues, we introduce an additio nal regularisation term which performs optimal transport in parallel within a lo w dimensional representation space. We demonstrate that operating in a low dimen sion representation of the data distribution benefits from convergence rate gain s in estimating the Wasserstein distance, resulting in more stable GAN training. We empirically show that our regulariser achieves a stabilising effect which le ads to higher quality of generated samples and increased mode coverage of the gi ven data distribution. Our method achieves significant improvements on the CIFAR -10, Oxford Flowers and CUB Birds datasets over several GAN baselines both quali tatively and quantitatively.

****

3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans

Ji Hou, Angela Dai, Matthias Niessner; Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4421-4430
We introduce 3D-SIS, a novel neural network architecture for 3D semantic instance segmentation in commodity RGB-D scans. The core idea of our method to jointly learn from both geometric and color signal, thus enabling accurate instance predictions. Rather than operate solely on 2D frames, we observe that most computer vision applications have multi-view RGB-D input available, which we leverage to construct an approach for 3D instance segmentation that effectively fuses together these multi-modal inputs. Our network leverages high-resolution RGB input by associating 2D images with the volumetric grid based on the pose alignment of the 3D reconstruction. For each image, we first extract 2D features for each pixel with a series of 2D convolutions; we then backproject the resulting feature vector to the associated voxel in the 3D grid. This combination of 2D and 3D feature learning allows significantly higher accuracy object detection and instance segmentation than state-of-the-art alternatives. We show results on both synthetic and real-world public benchmarks, achieving an improvement in mAP of over 13 on real-world data.
****

Causes and Corrections for Bimodal Multi-Path Scanning With Structured Light
Yu Zhang, Daniel L. Lau, Ying Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4431-4439
Structured light illumination is an active 3D scanning technique based on projecting/capturing a set of striped patterns and measuring the warping of the patterns as they reflect off a target object's surface. As designed, each pixel in the camera sees exactly one pixel from the projector; however, there are multi-path situations when the scanned surface has a complicated geometry with step edges and other discontinuities in depth or where the target surface has specularities that reflect light away from the camera. These situations are generally referred to multi-path where a camera pixel sees light from multiple projector positions. In the case of bimodal multi-path, the camera pixel receives light from exactly two positions which occurs along a step edge where the edge slices through a pixel so that the pixel sees both a foreground and background surface. In this paper, we present a general mathematical model to address the bimodal multi-path issue in a phase-measuring-profilometry scanner to measure the constructive and destructive interference between the two light paths, and by taking advantage of this interesting cue, separate the paths and make two decoupled phase measurements. We validate our algorithm with a number of challenging real-world scenarios, outperforming the state-of-the-art method.
****

TextureNet: Consistent Local Parametrizations for Learning From High-Resolution Signals on Meshes
Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Niessner, Leonidas J. Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4440-4449
We introduce, TextureNet, a neural network architecture designed to extract features from high-resolution signals associated with 3D surface meshes (e.g., color texture maps). The key idea is to utilize a 4-rotational symmetric(4-RoSy) field to define a domain for convolution on a surface. Though 4-RoSy fields have several properties favor-able for convolution on surfaces (low distortion, few singularities, consistent parameterization, etc.), orientations are ambiguous up to 4-fold rotation at any sample point. So, we introduce a new convolutional operator invariant to the4-RoSy ambiguity and use it in a network to extract features from high-resolution signals on geodesic neighborhoods of a surface. In comparison to alternatives, such as PointNet-based methods which lack a notion of orientation, the coherent structure given by these neighborhoods results in significantly stronger features. As an example application, we demonstrate the benefits of our architecture for 3D semantic segmentation of textured 3D meshes. The results show that our method outperforms all existing methods on the basis of mean IoU by a significant margin in both geometry-only(6.4%) and RGB+Geometry (6.9-8.2%) settings.
****

PlaneRCNN: 3D Plane Detection and Reconstruction From a Single Image
Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4450-4459
This paper proposes a deep neural architecture, PlaneRCNN, that detects and reconstructs piecewise planar regions from a single RGB image. PlaneRCNN employs a variant of Mask R-CNN to detect planes with their plane parameters and segmentation masks. PlaneRCNN then refines an arbitrary number of segmentation masks with a novel loss enforcing the consistency with a nearby view during training. The paper also presents a new benchmark with more fine-grained plane segmentations in the ground-truth, in which, PlaneRCNN outperforms existing state-of-the-art methods with significant margins in the plane detection, segmentation, and reconstruction metrics. PlaneRCNN makes an important step towards robust plane extraction method, which would have immediate impact on a wide range of applications including Robotics, Augmented Reality, and Virtual Reality.
****

Occupancy Networks: Learning 3D Reconstruction in Function Space
Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4460-4470
With the advent of deep neural networks, learning-based approaches for 3D reconstruction have gained popularity. However, unlike for images, in 3D there is no canonical representation which is both computationally and memory efficient yet allows for representing high-resolution geometry of arbitrary topology. Many of the state-of-the-art learning-based 3D reconstruction approaches can hence only represent very coarse 3D geometry or are limited to a restricted domain. In this paper, we propose Occupancy Networks, a new representation for learning-based 3D reconstruction methods. Occupancy networks implicitly represent the 3D surface as the continuous decision boundary of a deep neural network classifier. In contrast to existing approaches, our representation encodes a description of the 3D output at infinite resolution without excessive memory footprint. We validate that our representation can efficiently encode 3D structure and can be inferred from various kinds of input. Our experiments demonstrate competitive results, both qualitatively and quantitatively, for the challenging tasks of 3D reconstruction from single images, noisy point clouds and coarse discrete voxel grids. We believe that occupancy networks will become a useful tool in a wide variety of learning-based 3D tasks.
****

3D Shape Reconstruction From Images in the Frequency Domain
Weichao Shen, Yunde Jia, Yuwei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4471-4479
Reconstructing the high-resolution volumetric 3D shape from images is challenging due to the cubic growth of computational cost. In this paper, we propose a Fourier-based method that reconstructs a 3D shape from images in a 2D space by predicting slices in the frequency domain. According to the Fourier slice projection theorem, we introduce a thickness map to bridge the domain gap between images in the spatial domain and slices in the frequency domain. The thickness map is the 2D spatial projection of the 3D shape, which is easily predicted from the input image by a general convolutional neural network. Each slice in the frequency domain is the Fourier transform of the corresponding thickness map. All slices constitute a 3D descriptor and the 3D shape is the inverse Fourier transform of the descriptor. Using slices in the frequency domain, our method can transfer the 3D shape reconstruction from the 3D space into the 2D space, which significantly reduces the computational cost. The experiment results on the ShapeNet dataset demonstrate that our method achieves competitive reconstruction accuracy and computational efficiency compared with the state-of-the-art reconstruction methods.
****

SiCloPe: Silhouette-Based Clothed People
Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, Shigeo Morishima; Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR), 2019, pp. 4480-4490
We introduce a new silhouette-based representation for modeling clothed human bo dies using deep generative models. Our method can reconstruct a complete and tex tured 3D model of a person wearing clothes from a single input picture. Inspired by the visual hull algorithm, our implicit representation uses 2D silhouettes a nd 3D joints of a body pose to describe the immense shape complexity and variati ons of clothed people. Given a segmented 2D silhouette of a person and its infer red 3D joints from the input picture, we first synthesize consistent silhouettes from novel view points around the subject. The synthesized silhouettes which ar e the most consistent with the input segmentation are fed into a deep visual hul l algorithm for robust 3D shape prediction. We then infer the texture of the sub ject's back view using the frontal image and segmentation mask as input to a con ditional generative adversarial network. Our experiments demonstrate that our si lhouette-based model is an effective representation and the appearance of the ba ck view can be predicted reliably using an image-to-image translation network. W hile classic methods based on parametric models often fail for single-view image s of subjects with challenging clothing, our approach can still produce successf ul results, which are comparable to those obtained from multi-view input.
****

Detailed Human Shape Estimation From a Single Image by Hierarchical Mesh Deforma tion
Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, Ruigang Yang; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 44 91-4500
This paper presents a novel framework to recover detailed human body shapes from a single image. It is a challenging task due to factors such as variations in human shapes, body poses, and viewpoints. Prior methods typically attempt to re cover the human body shape using a parametric based template that lacks the surf ace details. As such the resulting body shape appears to be without clothing. In this paper, we propose a novel learning-based framework that combines the rob ustness of parametric model with the flexibility of free-form 3D deformation. W e use the deep neural networks to refine the 3D shape in a Hierarchical Mesh Def ormation (HMD) framework, utilizing the constraints from body joints, silhouette s, and per-pixel shading information. We are able to restore detailed human bod y shapes beyond skinned models. Experiments demonstrate that our method has out performed previous state-of-the-art approaches, achieving better accuracy in ter ms of both 2D IoU number and 3D metric distance. The code is available in https ://github.com/zhuhao-nju/hmd.git.
****

Convolutional Mesh Regression for Single-Image Human Shape Reconstruction
Nikos Kolotouros, Georgios Pavlakos, Kostas Daniilidis; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 45 01-4510
This paper addresses the problem of 3D human pose and shape estimation from a si ngle image. Previous approaches consider a parametric model of the human body, S MPL, and attempt to regress the model parameters that give rise to a mesh consis tent with image evidence. This parameter regression has been a very challenging task, with model-based approaches underperforming compared to nonparametric solu tions in terms of pose estimation. In our work, we propose to relax this heavy r eliance on the model's parameter space. We still retain the topology of the SMPL template mesh, but instead of predicting model parameters, we directly regress the 3D location of the mesh vertices. This is a heavy task for a typical network , but our key insight is that the regression becomes significantly easier using a Graph-CNN. This architecture allows us to explicitly encode the template mesh structure within the network and leverage the spatial locality the mesh has to o ffer. Image-based features are attached to the mesh vertices and the Graph-CNN i s responsible to process them on the mesh structure, while the regression target for each vertex is its 3D location. Having recovered the complete 3D geometry o f the mesh, if we still require a specific model parametrization, this can be re liably regressed from the vertices locations. We demonstrate the flexibility and

the effectiveness of our proposed graph-based mesh regression by attaching different types of features on the mesh vertices. In all cases, we outperform the comparable baselines relying on model parameter regression, while we also achieve state-of-the-art results among model-based pose estimation approaches.
****

H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions
Bugra Tekin, Federica Bogo, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4511-4520
We present a unified framework for understanding 3D hand and object interactions in raw image sequences from egocentric RGB cameras. Given a single RGB image, our model jointly estimates the 3D hand and object poses, models their interactions, and recognizes the object and action classes with a single feed-forward pass through a neural network. We propose a single architecture that does not rely on external detection algorithms but rather is trained end-to-end on single images. We further merge and propagate information in the temporal domain to infer interactions between hand and object trajectories and recognize actions. The complete model takes as input a sequence of frames and outputs per-frame 3D hand and object pose predictions along with the estimates of object and action categories for the entire sequence. We demonstrate state-of-the-art performance of our algorithm even in comparison to the approaches that work on depth data and ground-truth annotations.
****

Learning the Depths of Moving People by Watching Frozen People
Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, William T. Freeman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4521-4530
We present a method for predicting dense depth in scenarios where both a monocular camera and people in the scene are freely moving. Existing methods for recovering depth for dynamic, non-rigid objects from monocular video impose strong assumptions on the objects' motion and may only recover sparse depth. In this paper, we take a data-driven approach and learn human depth priors from a new source of data: thousands of Internet videos of people imitating mannequins, i.e., freezing in diverse, natural poses, while a hand-held camera tours the scene. Since the people are stationary, training data can be created from these videos using multi-view stereo reconstruction. At inference time, our method uses motion parallax cues from the static areas of the scenes, and shows clear improvement over state-of-the-art monocular depth prediction methods. We demonstrate our method on real-world sequences of complex human actions captured by a moving hand-held camera, and show various 3D effects produced using our predicted depth.
****

Extreme Relative Pose Estimation for RGB-D Scans via Scene Completion
Zhenpei Yang, Jeffrey Z. Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4531-4540
Estimating the relative rigid pose between two RGB-D scans of the same underlying environment is a fundamental problem in computer vision, robotics, and computer graphics. Most existing approaches allow only limited maximum relative pose changes since they require considerable overlap between the input scans. We introduce a novel approach that extends the scope to extreme relative poses, with little or even no overlap between the input scans. The key idea is to infer more complete scene information about the underlying environment and match on the completed scans. In particular, instead of only performing scene completion from each individual scan, our approach alternates between relative pose estimation and scene completion. This allows us to perform scene completion by utilizing information from both input scans at late iterations, resulting in better results for both scene completion and relative pose estimation. Experimental results on benchmark datasets show that our approach leads to considerable improvements over state-of-the-art approaches for relative pose estimation. In particular, our approach provides encouraging relative pose estimates even between non-overlapping scans.

****
A Skeleton-Bridged Deep Learning Approach for Generating Meshes of Complex Topologies From Single RGB Images
Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, Xin Tong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4541-4550
This paper focuses on the challenging task of learning 3D object surface reconstructions from single RGB images. Existing methods achieve varying degrees of success by using different geometric representations. However, they all have their own drawbacks, and cannot well reconstruct those surfaces of complex topologies. To this end, we propose in this paper a skeleton-bridged, stage-wise learning approach to address the challenge. Our use of skeleton is due to its nice property of topology preservation, while being of lower complexity to learn. To learn skeleton from an input image, we design a deep architecture whose decoder is based on a novel design of parallel streams respectively for synthesis of curve- and surface-like skeleton points. We use different shape representations of point cloud, volume, and mesh in our stage-wise learning, in order to take their respective advantages. We also propose multi-stage use of the input image to correct prediction errors that are possibly accumulated in each stage. We conduct intensive experiments to investigate the efficacy of our proposed approach. Qualitative and quantitative results on representative object categories of both simple and complex topologies demonstrate the superiority of our approach over existing ones. We will make our ShapeNet-Skeleton dataset publicly available.
****
Learning Structure-And-Motion-Aware Rolling Shutter Correction
Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4551-4560
An exact method of correcting the rolling shutter (RS) effect requires recovering the underlying geometry, i.e. the scene structures and the camera motions between scanlines or between views. However, the multiple-view geometry for RS cameras is much more complicated than its global shutter (GS) counterpart, with various degeneracies. In this paper, we first make a theoretical contribution by showing that RS two-view geometry is degenerate in the case of pure translational camera motion. In view of the complex RS geometry, we then propose a Convolutional Neural Network (CNN)-based method which learns the underlying geometry (camera motion and scene structure) from just a single RS image and perform RS image correction. We call our method structure-and-motion-aware RS correction because it reasons about the concealed motions between the scanlines as well as the scene structure. Our method learns from a large-scale dataset synthesized in a geometrically meaningful way where the RS effect is generated in a manner consistent with the camera motion and scene structure. In extensive experiments, our method achieves superior performance compared to other state-of-the-art methods for single image RS correction and subsequent Structure from Motion (SfM) applications.
****
PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation
Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, Hujun Bao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4561-4570
This paper addresses the challenge of 6DoF pose estimation from a single RGB image under severe occlusion or truncation. Many recent works have shown that a two-stage approach, which first detects keypoints and then solves a Perspective-n-Point (PnP) problem for pose estimation, achieves remarkable performance. However, most of these methods only localize a set of sparse keypoints by regressing their image coordinates or heatmaps, which are sensitive to occlusion and truncation. Instead, we introduce a Pixel-wise Voting Network (PVNet) to regress pixel-wise vectors pointing to the keypoints and use these vectors to vote for keypoint locations. This creates a flexible representation for localizing occluded or truncated keypoints. Another important feature of this representation is that it provides uncertainties of keypoint locations that can be further leveraged by the

PnP solver. Experiments show that the proposed approach outperforms the state of the art on the LINEMOD, Occlusion LINEMOD and YCB-Video datasets by a large margin, while being efficient for real-time pose estimation. We further create a Truncation LINEMOD dataset to validate the robustness of our approach against truncation. The code is available at https://zju3dv.github.io/pvnet/.
****

SelFlow: Self-Supervised Learning of Optical Flow
Pengpeng Liu, Michael Lyu, Irwin King, Jia Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4571-4580
We present a self-supervised learning approach for optical flow. Our method distills reliable flow estimations from non-occluded pixels, and uses these predictions as ground truth to learn optical flow for hallucinated occlusions. We further design a simple CNN to utilize temporal information from multiple frames for better flow estimation. These two principles lead to an approach that yields the best performance for unsupervised optical flow learning on the challenging benchmarks including MPI Sintel, KITTI 2012 and 2015. More notably, our self-supervised pre-trained model provides an excellent initialization for supervised fine-tuning. Our fine-tuned models achieve state-of-the-art results on all three datasets. At the time of writing, we achieve EPE=4.26 on the Sintel benchmark, outperforming all submitted methods.
****

Taking a Deeper Look at the Inverse Compositional Algorithm
Zhaoyang Lv, Frank Dellaert, James M. Rehg, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4581-4590
In this paper, we provide a modern synthesis of the classic inverse compositional algorithm for dense image alignment. We first discuss the assumptions made by this well-established technique, and subsequently propose to relax these assumptions by incorporating data-driven priors into this model. More specifically, we unroll a robust version of the inverse compositional algorithm and replace multiple components of this algorithm using more expressive models whose parameters we train in an end-to-end fashion from data. Our experiments on several challenging 3D rigid motion estimation tasks demonstrate the advantages of combining optimization with learning-based techniques, outperforming the classic inverse compositional algorithm as well as data-driven image-to-pose regression approaches.
****

Deeper and Wider Siamese Networks for Real-Time Visual Tracking
Zhipeng Zhang, Houwen Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4591-4600
Siamese networks have drawn great attention in visual tracking because of their balanced accuracy and speed. However, the backbone networks used in Siamese trackers are relatively shallow, such as AlexNet, which does not fully take advantage of the capability of modern deep neural networks. In this paper, we investigate how to leverage deeper and wider convolutional neural networks to enhance tracking robustness and accuracy. We observe that direct replacement of backbones with existing powerful architectures, such as ResNet and Inception, does not bring improvements. The main reasons are that 1) large increases in the receptive field of neurons lead to reduced feature discriminability and localization precision; and 2) the network padding for convolutions induces a positional bias in learning. To address these issues, we propose new residual modules to eliminate the negative impact of padding, and further design new architectures using these modules with controlled receptive field size and network stride. The designed architectures are lightweight and guarantee real-time tracking speed when applied to SiamFC and SiamRPN. Experiments show that solely due to the proposed network architectures, our SiamFC+ and SiamRPN+ obtain up to 9.8%/6.3% (AUC), 23.3%/8.8% (EAO) and 24.4%/25.0% (EAO) relative improvements over the original versions on the OTB-15, VOT-16 and VOT-17 datasets, respectively.
****

Self-Supervised Adaptation of High-Fidelity Face Models for Monocular Performance Tracking

Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, Hyun Soo Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4601-4609

Improvements in data-capture and face modeling techniques have enabled us to create high-fidelity realistic face models. However, driving these realistic face models requires special input data, e.g., 3D meshes and unwrapped textures. Also, these face models expect clean input data taken under controlled lab environments, which is very different from data collected in the wild. All these constraints make it challenging to use the high-fidelity models in tracking for commodity cameras. In this paper, we propose a self-supervised domain adaptation approach to enable the animation of high-fidelity face models from a commodity camera. Our approach first circumvents the requirement for special input data by training a new network that can directly drive a face model just from a single 2D image. Then, we overcome the domain mismatch between lab and uncontrolled environments by performing self-supervised domain adaptation based on "consecutive frame texture consistency" based on the assumption that the appearance of the face is consistent over consecutive frames, avoiding the necessity of modeling the new environment such as lighting or background. Experiments show that we are able to drive a high-fidelity face model to perform complex facial motion from a cellphone camera without requiring any labeled data from the new domain.

****

Diverse Generation for Multi-Agent Sports Games

Raymond A. Yeh, Alexander G. Schwing, Jonathan Huang, Kevin Murphy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4610-4619

In this paper, we propose a new generative model for multi-agent trajectory data, focusing on the case of multi-player sports games. Our model leverages graph neural networks (GNNs) and variational recurrent neural networks (VRNNs) to achieve a permutation equivariant model suitable for sports. On two challenging datasets (basketball and soccer), we show that we are able to produce more accurate forecasts than previous methods. We assess accuracy using various metrics, such as log-likelihood and "best of N" loss, based on N different samples of the future. We also measure the distribution of statistics of interest, such as player location or velocity, and show that the distribution induced by our generative model better matches the empirical distribution of the test set. Finally, we show that our model can perform conditional prediction, which lets us answer counterfactual questions such as "how will the players move differently if A passes the ball to B instead of C?"

****

Efficient Online Multi-Person 2D Pose Tracking With Recurrent Spatio-Temporal Affinity Fields

Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, Yaser Sheikh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4620-4628

We present an online approach to efficiently and simultaneously detect and track 2D poses of multiple people in a video sequence. We build upon Part Affinity Field (PAF) representation designed for static images, and propose an architecture that can encode and predict Spatio-Temporal Affinity Fields (STAF) across a video sequence. In particular, we propose a novel temporal topology cross-linked across limbs which can consistently handle body motions of a wide range of magnitudes. Additionally, we make the overall approach recurrent in nature, where the network ingests STAF heatmaps from previous frames and estimates those for the current frame. Our approach uses only online inference and tracking, and is currently the fastest and the most accurate bottom-up approach that is runtime-invariant to the number of people in the scene and accuracy-invariant to input frame rate of camera. Running at ~30 fps on a single GPU at single scale, it achieves highly competitive results on the PoseTrack benchmarks.

****

GFrames: Gradient-Based Local Reference Frame for 3D Shape Matching

Simone Melzi, Riccardo Spezialetti, Federico Tombari, Michael M. Bronstein,

Luigi Di Stefano, Emanuele Rodola; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4629-4638
We introduce GFrames, a novel local reference frame (LRF) construction for 3D meshes and point clouds. GFrames are based on the computation of the intrinsic gradient of a scalar field defined on top of the input shape. The resulting tangent vector field defines a repeatable tangent direction of the local frame at each point; importantly, it directly inherits the properties and invariance classes of the underlying scalar function, making it remarkably robust under strong sampling artifacts, vertex noise, as well as non-rigid deformations. Existing local descriptors can directly benefit from our repeatable frames, as we showcase in a selection of 3D vision and shape analysis applications where we demonstrate state-of-the-art performance in a variety of challenging settings.
****

Eliminating Exposure Bias and Metric Mismatch in Multiple Object Tracking
Andrii Maksai, Pascal Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4639-4648
Identity Switching remains one of the main difficulties Multiple Object Tracking (MOT) algorithms have to deal with. Many state-of-the-art approaches now use sequence models to solve this problem but their training can be affected by biases that decrease their efficiency. In this paper, we introduce a new training procedure that confronts the algorithm to its own mistakes while explicitly attempting to minimize the number of switches, which results in better training. We propose an iterative scheme of building a rich training set and using it to learn a scoring function that is an explicit proxy for the target tracking metric. Whether using only simple geometric features or more sophisticated ones that also take appearance into account, our approach outperforms the state-of-the-art on several MOT benchmarks.
****

Graph Convolutional Tracking
Junyu Gao, Tianzhu Zhang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4649-4659
Tracking by siamese networks has achieved favorable performance in recent years. However, most of existing siamese methods do not take full advantage of spatial-temporal target appearance modeling under different contextual situations. In fact, the spatial-temporal information can provide diverse features to enhance the target representation, and the context information is important for online adaption of target localization. To comprehensively leverage the spatial-temporal structure of historical target exemplars and get benefit from the context information, in this work, we present a novel Graph Convolutional Tracking (GCT) method for high-performance visual tracking. Specifically, the GCT jointly incorporates two types of Graph Convolutional Networks (GCNs) into a siamese framework for target appearance modeling. Here, we adopt a spatial-temporal GCN to model the structured representation of historical target exemplars. Furthermore, a context GCN is designed to utilize the context of the current frame to learn adaptive features for target localization. Extensive results on 4 challenging benchmarks show that our GCT method performs favorably against state-of-the-art trackers while running around 50 frames per second.
****

ATOM: Accurate Tracking by Overlap Maximization
Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, Michael Felsberg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4660-4669
While recent years have witnessed astonishing improvements in visual tracking robustness, the advancements in tracking accuracy have been limited. As the focus has been directed towards the development of powerful classifiers, the problem of accurate target state estimation has been largely overlooked. In fact, most trackers resort to a simple multi-scale search in order to estimate the target bounding box. We argue that this approach is fundamentally limited since target estimation is a complex task, requiring high-level knowledge about the object. We address this problem by proposing a novel tracking architecture, consisting of d

edicated target estimation and classification components. High level knowledge i s incorporated into the target estimation through extensive offline learning. Ou r target estimation component is trained to predict the overlap between the targ et object and an estimated bounding box. By carefully integrating target-specifi c information, our approach achieves previously unseen bounding box accuracy. We further introduce a classification component that is trained online to guarante e high discriminative power in the presence of distractors. Our final tracking f ramework sets a new state-of-the-art on five challenging benchmarks. On the new large-scale TrackingNet dataset, our tracker ATOM achieves a relative gain of 15 % over the previous best approach, while running at over 30 FPS. Code and models are available at https://github.com/visionml/pytracking.
****

Visual Tracking via Adaptive Spatially-Regularized Correlation Filters
Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, Jianhua Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4670-4679
In this work, we propose a novel adaptive spatially-regularized correlation filt ers (ASRCF) model to simultaneously optimize the filter coefficients and the spa tial regularization weight. First, this adaptive spatial regularization scheme c ould learn an effective spatial weight for a specific object and its appearance variations, and therefore result in more reliable filter coefficients during the tracking process. Second, our ASRCF model can be effectively optimized based on the alternating direction method of multipliers, where each subproblem has the closed-from solution. Third, our tracker applies two kinds of CF models to estim ate the location and scale respectively. The location CF model exploits ensemble s of shallow and deep features to determine the optimal position accurately. The scale CF model works on multi-scale shallow features to estimate the optimal sc ale efficiently. Extensive experiments on five recent benchmarks show that our t racker performs favorably against many state-of-the-art algorithms, with real-ti me performance of 28fps.
****

Deep Tree Learning for Zero-Shot Face Anti-Spoofing
Yaojie Liu, Joel Stehouwer, Amin Jourabloo, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4680-4689
Face anti-spoofing is designed to keep face recognition systems from recognizing fake faces as the genuine users. While advanced face anti-spoofing methods are developed, new types of spoof attacks are also being created and becoming a thre at to all existing systems. We define the detection of unknown spoof attacks as Zero-Shot Face Anti-spoofing (ZSFA). Previous works of ZSFA only study 1-2 types of spoof attacks, such as print/replay attacks, which limits the insight of thi s problem. In this work, we expand the ZSFA problem to a wide range of 13 types of spoof attacks, including print attack, replay attack, 3D mask attacks, and so on. A novel Deep Tree Network (DTN) is proposed to tackle the ZSFA. The tree is learned to partition the spoof samples into semantic sub-groups in an unsupervi sed fashion. When a data sample arrives, being know or unknown attacks, DTN rout es it to the most similar spoof cluster, and make the binary decision. In additi on, to enable the study of ZSFA, we introduce the first face anti-spoofing datab ase that contains diverse types of spoof attacks. Experiments show that our prop osed method achieves the state of the art on multiple testing protocols of ZSFA.
****

ArcFace: Additive Angular Margin Loss for Deep Face Recognition
Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690-4699
One of the main challenges in feature learning using Deep Convolutional Neural N etworks (DCNNs) for large-scale face recognition is the design of appropriate lo ss functions that can enhance the discriminative power. Centre loss penalises th e distance between deep features and their corresponding class centres in the Eu clidean space to achieve intra-class compactness. SphereFace assumes that the li

near transformation matrix in the last fully connected layer can be used as a re
presentation of the class centres in the angular space and therefore penalises t
he angles between deep features and their corresponding weights in a multiplicat
ive way. Recently, a popular line of research is to incorporate margins in well-
established loss functions in order to maximise face class separability. In this
 paper, we propose an Additive Angular Margin Loss (ArcFace) to obtain highly di
scriminative features for face recognition. The proposed ArcFace has a clear geo
metric interpretation due to its exact correspondence to geodesic distance on a
hypersphere. We present arguably the most extensive experimental evaluation agai
nst all recent state-of-the-art face recognition methods on ten face recognition
 benchmarks which includes a new large-scale image database with trillions of pa
irs and a large-scale video dataset. We show that ArcFace consistently outperfor
ms the state of the art and can be easily implemented with negligible computatio
nal overhead. To facilitate future research, the code has been made available.
****

Learning Joint Gait Representation via Quintuplet Loss Minimization
Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, Hongdong Li; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.
4700-4709
Gait recognition is an important biometric method popularly used in video survei
llance, where the task is to identify people at a distance by their walking patt
erns from video sequences. Most of the current successful approaches for gait re
cognition either use a pair of gait images to form a cross-gait representation o
r rely on a single gait image for unique-gait representation. These two types of
 representations emperically complement one another. In this paper, we propose a
 new Joint Unique-gait and Cross-gait Network (JUCNet), to combine the advantage
s of unique-gait representation with that of cross-gait representation, leading
to an significantly improved performance. Another key contribution of this paper
 is a novel quintuplet loss function, which simultaneously increases the inter-c
lass differences by pushing representations extracted from different subjects ap
art and decreases the intra-class variations by pulling representations extracte
d from the same subject together. Experiments show that our method achieves the
state-of-the-art performance tested on standard benchmark datasets, demonstratin
g its superiority over existing methods.
****

Gait Recognition via Disentangled Representation Learning
Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, Nan
xin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2019, pp. 4710-4719
Gait, the walking pattern of individuals, is one of the most important biometric
s modalities. Most of the existing gait recognition methods take silhouettes or
articulated body models as the gait features. These methods suffer from degraded
 recognition performance when handling confounding variables, such as clothing,
carrying and view angle. To remedy this issue, we propose a novel AutoEncoder fr
amework to explicitly disentangle pose and appearance features from RGB imagery
and the LSTM-based integration of pose features over time produces the gait feat
ure. In addition, we collect a Frontal-View Gait (FVG) dataset to focus on gait
recognition from frontal-view walking, which is a challenging problem since it c
ontains minimal gait cues compared to other views. FVG also includes other impor
tant variations,e.g., walking speed, carrying, and clothing. With extensive expe
riments on CASIA-B, USF and FVG datasets, our method demonstrates superior perfo
rmance to the-state-of-the-arts quantitatively, the ability of feature disentang
lement qualitatively, and promising computational efficiency.
****

Reversible GANs for Memory-Efficient Image-To-Image Translation
Tycho F.A. van der Ouderaa, Daniel E. Worrall; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4720-4728
The pix2pix and CycleGAN losses have vastly improved the qualitative and quantit
ative visual quality of results in image-to-image translation tasks. We extend t
his framework by exploring approximately invertible architectures which are well

suited to these losses. These architectures are approximately invertible by des
ign and thus partially satisfy cycle-consistency before training even begins. Fu
rthermore, since invertible architectures have constant memory complexity in dep
th, these models can be built arbitrarily deep. We are able to demonstrate super
ior quantitative output on the Cityscapes and Maps datasets at near constant mem
ory budget.
****

Sensitive-Sample Fingerprinting of Deep Neural Networks
Zecheng He, Tianwei Zhang, Ruby Lee; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4729-4737
Numerous cloud-based services are provided to help customers develop and deploy
deep learning applications. When a customer deploys a deep learning model in the
cloud and serves it to end-users, it is important to be able to verify that the
deployed model has not been tampered with. In this paper, we propose a novel a
nd practical methodology to verify the integrity of remote deep learning models,
with only black-box access to the target models. Specifically, we define Sensit
ive-Sample fingerprints, which are a small set of human unnoticeable transformed
inputs that make the model outputs sensitive to the model's parameters. Even sm
all model changes can be clearly reflected in the model outputs. Experimental re
sults on different types of model integrity attacks show that we proposed approa
ch is both effective and efficient. It can detect model integrity breaches with
high accuracy (>99.95%) and guaranteed zero false positives on all evaluated att
acks. Meanwhile, it only requires up to 103X fewer model inferences, compared wi
th non-sensitive samples.
****

Soft Labels for Ordinal Regression
Raul Diaz, Amit Marathe; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2019, pp. 4738-4747
Ordinal regression attempts to solve classification problems in which categories
are not independent, but rather follow a natural order. It is crucial to classi
fy each class correctly while learning adequate interclass ordinal relationships
. We present a simple and effective method that constrains these relationships a
mong categories by seamlessly incorporating metric penalties into ground truth l
abel representations. This encoding allows deep neural networks to automatically
learn intraclass and interclass relationships without any explicit modification
of the network architecture. Our method converts data labels into soft probabil
ity distributions that pair well with common categorical loss functions such as
cross-entropy. We show that this approach is effective by using off-the-shelf cl
assification and segmentation networks in four wildly different scenarios: image
quality ranking, age estimation, horizon line regression, and monocular depth e
stimation. We demonstrate that our general-purpose method is very competitive wi
th respect to specialized approaches, and adapts well to a variety of different
network architectures and metrics.
****

Local to Global Learning: Gradually Adding Classes for Training Deep Neural Netw
orks
Hao Cheng, Dongze Lian, Bowen Deng, Shenghua Gao, Tao Tan, Yanlin Geng; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2019, pp. 4748-4756
We propose a new learning paradigm, Local to Global Learning (LGL), for Deep Neu
ral Networks (DNNs) to improve the performance of classification problems. The c
ore of LGL is to learn a DNN model from fewer categories (local) to more categor
ies (global) gradually within the entire training set. LGL is most related to th
e Self-Paced Learning (SPL) algorithm but its formulation is different from SPL.
SPL trains its data from simple to complex, while LGL from local to global. In
this paper, we incorporate the idea of LGL into the learning objective of DNNs a
nd explain why LGL works better from an information-theoretic perspective. Exper
iments on the toy data, CIFAR-10, CIFAR-100, and ImageNet dataset show that LGL
outperforms the baseline and SPL-based algorithms.
****

What Does It Mean to Learn in Deep Networks? And, How Does One Detect Adversaria l Attacks?
Ciprian A. Corneanu, Meysam Madadi, Sergio Escalera, Aleix M. Martinez; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2019, pp. 4757-4766
The flexibility and high-accuracy of Deep Neural Networks (DNNs) has transformed computer vision. But, the fact that we do not know when a specific DNN will wor k and when it will fail has resulted in a lack of trust. A clear example is self -driving cars; people are uncomfortable sitting in a car driven by algorithms th at may fail under some unknown, unpredictable conditions. Interpretability and e xplainability approaches attempt to address this by uncovering what a DNN models , i.e., what each node (cell) in the network represents and what images are most likely to activate it. This can be used to generate, for example, adversarial a ttacks. But these approaches do not generally allow us to determine where a DNN will succeed or fail and why . i.e., does this learned representation genera lize to unseen samples? Here, we derive a novel approach to define what it mean s to learn in deep networks, and how to use this knowledge to detect adversarial attacks. We show how this defines the ability of a network to generalize to uns een testing samples and, most importantly, why this is the case.
****

Handwriting Recognition in Low-Resource Scripts Using Adversarial Learning
Ayan Kumar Bhunia, Abhirup Das, Ankan Kumar Bhunia, Perla Sai Raj Kishore, P artha Pratim Roy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4767-4776
Handwritten Word Recognition and Spotting is a challenging field dealing with ha ndwritten text possessing irregular and complex shapes. The design of deep neura l network models makes it necessary to extend training datasets in order to intr oduce variations and increase the number of samples; word-retrieval is therefore very difficult in low-resource scripts. Much of the existing literature compris es preprocessing strategies which are seldom sufficient to cover all possible va riations. We propose an Adversarial Feature Deformation Module (AFDM) that learn s ways to elastically warp extracted features in a scalable manner. The AFDM is inserted between intermediate layers and trained alternatively with the original framework, boosting its capability to better learn highly informative features rather than trivial ones. We test our meta-framework, which is built on top of p opular word-spotting and word-recognition frameworks and enhanced by AFDM, not o nly on extensive Latin word datasets but also on sparser Indic scripts. We recor d results for varying sizes of training data, and observe that our enhanced netw ork generalizes much better in the low-data regime; the overall word-error rates and mAP scores are observed to improve as well.
****

Adversarial Defense Through Network Profiling Based Path Extraction
Yuxian Qiu, Jingwen Leng, Cong Guo, Quan Chen, Chao Li, Minyi Guo, Yuhao Z hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2019, pp. 4777-4786
Recently, researchers have started decomposing deep neural network models accord ing to their semantics or functions. Recent work has shown the effectiveness of decomposed functional blocks for defending adversarial attacks, which add small input perturbation to the input image to fool the DNN models. This work proposes a profiling-based method to decompose the DNN models to different functional bl ocks, which lead to the effective path as a new approach to exploring DNNs' inte rnal organization. Specifically, the per-image effective path can be aggregated to the class-level effective path, through which we observe that adversarial ima ges activate effective path different from normal images. We propose an effectiv e path similarity-based method to detect adversarial images with an interpretabl e model, which achieve better accuracy and broader applicability than the state- of-the-art technique.
****

RENAS: Reinforced Evolutionary Neural Architecture Search
Yukang Chen, Gaofeng Meng, Qian Zhang, Shiming Xiang, Chang Huang, Lisen Mu

, Xinggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4787-4796
Neural Architecture Search (NAS) is an important yet challenging task in network design due to its high computational consumption. To address this issue, we propose the Reinforced Evolutionary Neural Architecture Search (RENAS), which is an evolutionary method with reinforced mutation for NAS. Our method integrates reinforced mutation into an evolution algorithm for neural architecture exploration, in which a mutation controller is introduced to learn the effects of slight modifications and make mutation actions. The reinforced mutation controller guides the model population to evolve efficiently. Furthermore, as child models can inherit parameters from their parents during evolution, our method requires very limited computational resources. In experiments, we conduct the proposed search method on CIFAR-10 and obtain a powerful network architecture, RENASNet. This architecture achieves a competitive result on CIFAR-10. The explored network architecture is transferable to ImageNet and achieves a new state-of-the-art accuracy, i.e., 75.7% top-1 accuracy with 5.36M parameters on mobile ImageNet. We further test its performance on semantic segmentation with DeepLabv3 on the PASCAL VOC. RENASNet outperforms MobileNet-v1, MobileNet-v2 and NASNet. It achieves 75.83% mIOU without being pretrained on COCO.
****

Co-Occurrence Neural Network
Irina Shevlev, Shai Avidan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4797-4804
Convolutional Neural Networks (CNNs) became a very popular tool for image analysis. Convolutions are fast to compute and easy to store, but they also have some limitations. First, they are shift-invariant and, as a result, they do not adapt to different regions of the image. Second, they have a fixed spatial layout, so small geometric deformations in the layout of a patch will completely change the filter response. For these reasons, we need multiple filters to handle the different parts and variations in the input. We augment the standard convolutional tools used in CNNs with a new filter that addresses both issues raised above. Our filter combines two terms, a spatial filter and a term that is based on the co-occurrence statistics of input values in the neighborhood. The proposed filter is differentiable and can therefore be packaged as a layer in CNN and trained using back-propagation. We show how to train the filter as part of the network and report results on several data sets. In particular, we replace a convolutional layer with hundreds of thousands of parameters with a Co-occurrence Layer consisting of only a few hundred parameters with minimal impact on accuracy.
****

SpotTune: Transfer Learning Through Adaptive Fine-Tuning
Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, Rogerio Feris; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4805-4814
Transfer learning, which allows a source task to affect the inductive bias of the target task, is widely used in computer vision. The typical way of conducting transfer learning with deep neural networks is to fine-tune a model pretrained on the source task using data from the target task. In this paper, we propose an adaptive fine-tuning approach, called SpotTune, which finds the optimal fine-tuning strategy per instance for the target data. In SpotTune, given an image from the target task, a policy network is used to make routing decisions on whether to pass the image through the fine-tuned layers or the pre-trained layers. We conduct extensive experiments to demonstrate the effectiveness of the proposed approach. Our method outperforms the traditional fine-tuning approach on 12 out of 14 standard datasets. We also compare SpotTune with other state-of-the-art fine-tuning strategies, showing superior performance. On the Visual Decathlon datasets, our method achieves the highest score across the board without bells and whistles.
****

Signal-To-Noise Ratio: A Robust Distance Metric for Deep Metric Learning
Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, Binghui Chen; Proceeding

s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4815-4824
Deep metric learning, which learns discriminative features to process image clustering and retrieval tasks, has attracted extensive attention in recent years. A number of deep metric learning methods, which ensure that similar examples are mapped close to each other and dissimilar examples are mapped farther apart, have been proposed to construct effective structures for loss functions and have shown promising results. In this paper, different from the approaches on learning the loss structures, we propose a robust SNR distance metric based on Signal-to-Noise Ratio (SNR) for measuring the similarity of image pairs for deep metric learning. By exploring the properties of our SNR distance metric from the view of geometry space and statistical theory, we analyze the properties of our metric and show that it can preserve the semantic similarity between image pairs, which well justify its suitability for deep metric learning. Compared with Euclidean distance metric, our SNR distance metric can further jointly reduce the intra-class distances and enlarge the inter-class distances for learned features. Leveraging our SNR distance metric, we propose Deep SNR-based Metric Learning (DSML) to generate discriminative feature embeddings. By extensive experiments on three widely adopted benchmarks, including CARS196, CUB200-2011 and CIFAR10, our DSML has shown its superiority over other state-of-the-art methods. Additionally, we extend our SNR distance metric to deep hashing learning, and conduct experiments on two benchmarks, including CIFAR10 and NUS-WIDE, to demonstrate the effectiveness and generality of our SNR distance metric.
****

Detection Based Defense Against Adversarial Examples From the Steganalysis Point of View
Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4825-4834
Deep Neural Networks (DNNs) have recently led to significant improvements in many fields. However, DNNs are vulnerable to adversarial examples which are samples with imperceptible perturbations while dramatically misleading the DNNs. Moreover, adversarial examples can be used to perform an attack on various kinds of DNN based systems, even if the adversary has no access to the underlying model. Many defense methods have been proposed, such as obfuscating gradients of the networks or detecting adversarial examples. However it is proved out that these defense methods are not effective or cannot resist secondary adversarial attacks. In this paper, we point out that steganalysis can be applied to adversarial examples detection, and propose a method to enhance steganalysis features by estimating the probability of modifications caused by adversarial attacks. Experimental results show that the proposed method can accurately detect adversarial examples. Moreover, secondary adversarial attacks are hard to be directly performed to our method because our method is not based on a neural network but based on high-dimensional artificial features and Fisher Linear Discriminant ensemble.
****

HetConv: Heterogeneous Kernel-Based Convolutions for Deep CNNs
Pravendra Singh, Vinay Kumar Verma, Piyush Rai, Vinay P. Namboodiri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4835-4844
We present a novel deep learning architecture in which the convolution operation leverages heterogeneous kernels. The proposed HetConv (Heterogeneous Kernel-Based Convolution) reduces the computation (FLOPs) and the number of parameters as compared to standard convolution operation while still maintaining representational efficiency. To show the effectiveness of our proposed convolution, we present extensive experimental results on the standard convolutional neural network (CNN) architectures such as VGG and ResNet. We find that after replacing the standard convolutional filters in these architectures with our proposed HetConv filters, we achieve 3X to 8X FLOPs based improvement in speed while still maintaining (and sometimes improving) the accuracy. We also compare our proposed convolutions with group/depth wise convolutions and show that it achieves more FLOPs redu

ction with significantly higher accuracy.
****
Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Fami liar Objects
Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn K u, Anh Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2019, pp. 4845-4854
Despite excellent performance on stationary test sets, deep neural networks (DNN s) can fail to generalize to out-of-distribution (OoD) inputs, including natural , non-adversarial ones, which are common in real-world settings. In this paper, we present a framework for discovering DNN failures that harnesses 3D renderers and 3D models. That is, we estimate the parameters of a 3D renderer that cause a target DNN to misbehave in response to the rendered image. Using our framework and a self-assembled dataset of 3D objects, we investigate the vulnerability of DNNs to OoD poses of well-known objects in ImageNet. For objects that are readil y recognized by DNNs in their canonical poses, DNNs incorrectly classify 97% of their pose space. In addition, DNNs are highly sensitive to slight pose perturba tions. Importantly, adversarial poses transfer across models and datasets. We fi nd that 99.9% and 99.4% of the poses misclassified by Inception-v3 also transfer to the AlexNet and ResNet-50 image classifiers trained on the same ImageNet dat aset, respectively, and 75.5% transfer to the YOLOv3 object detector trained on MS COCO.
****
Blind Geometric Distortion Correction on Images Through Deep Learning
Xiaoyu Li, Bo Zhang, Pedro V. Sander, Jing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4855-486 4
We propose the first general framework to automatically correct different types of geometric distortion in a single input image. Our proposed method employs con volutional neural networks (CNNs) trained by using a large synthetic distortion dataset to predict the displacement field between distorted images and corrected images. A model fitting method uses the CNN output to estimate the distortion p arameters, achieving a more accurate prediction. The final corrected image is ge nerated based on the predicted flow using an efficient, high-quality resampling method. Experimental results demonstrate that our algorithm outperforms traditio nal correction methods, and allows for interesting applications such as distorti on transfer, distortion exaggeration, and co-occurring distortion correction.
****
Instance-Level Meta Normalization
Songhao Jia, Ding-Jie Chen, Hwann-Tzong Chen; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4865-4873
This paper presents a normalization mechanism called Instance-Level Meta Normali zation (ILM Norm) to address a learning-to-normalize problem. ILM Norm learns to predict the normalization parameters via both the feature feed-forward and the gradient back-propagation paths. ILM Norm provides a meta normalization mechanis m and has several good properties. It can be easily plugged into existing instan ce-level normalization schemes such as Instance Normalization, Layer Normalizati on, or Group Normalization. ILM Norm normalizes each instance individually and t herefore maintains high performance even when small mini-batch is used. The expe rimental results show that ILM Norm well adapts to different network architectur es and tasks, and it consistently improves the performance of the original model s.
****
Iterative Normalization: Beyond Standardization Towards Efficient Whitening
Lei Huang, Yi Zhou, Fan Zhu, Li Liu, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4874-488 3
Batch Normalization (BN) is ubiquitously employed for accelerating neural networ k training and improving the generalization capability by performing standardiza tion within mini-batches. Decorrelated Batch Normalization (DBN) further boosts

the above effectiveness by whitening. However, DBN relies heavily on either a large batch size, or eigen-decomposition that suffers from poor efficiency on GPUs. We propose Iterative Normalization (IterNorm), which employs Newton's iterations for much more efficient whitening, while simultaneously avoiding the eigen-decomposition. Furthermore, we develop a comprehensive study to show IterNorm has better trade-off between optimization and generalization, with theoretical and experimental support. To this end, we exclusively introduce Stochastic Normalization Disturbance (SND), which measures the inherent stochastic uncertainty of samples when applied to normalization operations. With the support of SND, we provide natural explanations to several phenomena from the perspective of optimization, e.g., why group-wise whitening of DBN generally outperforms full-whitening and why the accuracy of BN degenerates with reduced batch sizes. We demonstrate the consistently improved performance of IterNorm with extensive experiments on CIFAR-10 and ImageNet over BN and DBN.
****

On Learning Density Aware Embeddings
Soumyadeep Ghosh, Richa Singh, Mayank Vatsa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4884-4892
Deep metric learning algorithms have been utilized to learn discriminative and generalizable models which are effective for classifying unseen classes. In this paper, a novel noise tolerant deep metric learning algorithm is proposed. The proposed method, termed as Density Aware Metric Learning, enforces the model to learn embeddings that are pulled towards the most dense region of the clusters for each class. It is achieved by iteratively shifting the estimate of the center towards the dense region of the cluster thereby leading to faster convergence and higher generalizability. In addition to this, the approach is robust to noisy samples in the training data, often present as outliers. Detailed experiments and analysis on two challenging cross-modal face recognition databases and two popular object recognition databases exhibit the efficacy of the proposed approach. It has superior convergence, requires lesser training time, and yields better accuracies than several popular deep metric learning methods.
****

Contrastive Adaptation Network for Unsupervised Domain Adaptation
Guoliang Kang, Lu Jiang, Yi Yang, Alexander G. Hauptmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4893-4902
Unsupervised Domain Adaptation (UDA) makes predictions for the target domain data while manual annotations are only available in the source domain. Previous methods minimize the domain discrepancy neglecting the class information, which may lead to misalignment and poor generalization performance. To address this issue, this paper proposes Contrastive Adaptation Network (CAN) optimizing a new metric which explicitly models the intra-class domain discrepancy and the inter-class domain discrepancy. We design an alternating update strategy for training CAN in an end-to-end manner. Experiments on two real-world benchmarks Office-31 and VisDA-2017 demonstrate that CAN performs favorably against the state-of-the-art methods and produces more discriminative features.
****

LP-3DCNN: Unveiling Local Phase in 3D Convolutional Neural Networks
Sudhakar Kumawat, Shanmuganathan Raman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4903-4912
Traditional 3D Convolutional Neural Networks (CNNs) are computationally expensive, memory intensive, prone to overfit, and most importantly, there is a need to improve their feature learning capabilities. To address these issues, we propose Rectified Local Phase Volume (ReLPV) block, an efficient alternative to the standard 3D convolutional layer. The ReLPV block extracts the phase in a 3D local neighborhood (e.g., 3x3x3) of each position of the input map to obtain the feature maps. The phase is extracted by computing 3D Short Term Fourier Transform (STFT) at multiple fixed low frequency points in the 3D local neighborhood of each position. These feature maps at different frequency points are then linearly combined after passing them through an activation function. The ReLPV block provide

s significant parameter savings of at least, 3^3 to 13^3 times compared to the s tandard 3D convolutional layer with the filter sizes 3x3x3 to 13x13x13, respecti vely. We show that the feature learning capabilities of the ReLPV block are sig nificantly better than the standard 3D convolutional layer. Furthermore, it prod uces consistently better results across different 3D data representations. We ac hieve state-of-the-art accuracy on the volumetric ModelNet10 and ModelNet40 data sets while utilizing only 11% parameters of the current state-of-the-art. We als o improve the state-of-the-art on the UCF-101 split-1 action recognition datase t by 5.68% (when trained from scratch) while using only 15% of the parameters of the state-of-the-art.
****

Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification
Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, Xian-sheng Hua; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 4913-4922
Video-based person re-identification plays an important role in surveillance vid eo analysis, expanding image-based methods by learning features of multiple fram es. Most existing methods fuse features by temporal average-pooling, without exp loring the different frame weights caused by various viewpoints, poses, and occl usions. In this paper, we propose an attribute-driven method for feature disenta ngling and frame re-weighting. The features of single frames are disentangled in to groups of sub-features, each corresponds to specific semantic attributes. The sub-features are re-weighted by the confidence of attribute recognition and the n aggregated at the temporal dimension as the final representation. By means of this strategy, the most informative regions of each frame are enhanced and contr ibutes to a more discriminative sequence representation. Extensive ablation stud ies demonstrate the effectiveness of feature disentangling as well as temporal r e-weighting. The experimental results on the iLIDS-VID, PRID-2011 and MARS datas ets demonstrate that our proposed method outperforms existing state-of-the-art a pproaches.
****

Binary Ensemble Neural Network: More Bits per Network or More Networks per Bit?
Shilin Zhu, Xin Dong, Hao Su; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2019, pp. 4923-4932
Binary neural networks (BNN) have been studied extensively since they run dramat ically faster at lower memory and power consumption than floating-point networks , thanks to the efficiency of bit operations. However, contemporary BNNs whose w eights and activations are both single bits suffer from severe accuracy degradat ion. To understand why, we investigate the representation ability, speed and bia s/variance of BNNs through extensive experiments. We conclude that the error of BNNs is predominantly caused by intrinsic instability (training time) and non-ro bustness (train & test time). Inspired by this investigation, we propose the Bin ary Ensemble Neural Network (BENN) which leverages ensemble methods to improve t he performance of BNNs with limited efficiency cost. While ensemble techniques h ave been broadly believed to be only marginally helpful for strong classifiers s uch as deep neural networks, our analysis and experiments show that they are nat urally a perfect fit to boost BNNs. We find that our BENN, which is faster and m ore robust than state-of-the-art binary networks, can even surpass the accuracy of the full-precision floating number network with the same architecture.
****

Distilling Object Detectors With Fine-Grained Feature Imitation
Tao Wang, Li Yuan, Xiaopeng Zhang, Jiashi Feng; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4933-4942
State-of-the-art CNN based recognition models are often computationally prohibit ive to deploy on low-end devices. A promising high level approach tackling this limitation is knowledge distillation, which let small student model mimic cumber some teacher model's output to get improved generalization. However, related met hods mainly focus on simple task of classification while do not consider complex tasks like object detection. We show applying the vanilla knowledge distillatio

n to detection model gets minor gain. To address the challenge of distilling kn owledge in detection model, we propose a fine-grained feature imitation method e xploiting the cross-location discrepancy of feature response. Our intuition is t hat detectors care more about local near object regions. Thus the discrepancy of feature response on the near object anchor locations reveals important informat ion of how teacher model tends to generalize. We design a novel mechanism to est imate those locations and let student model imitate the teacher on them to get e nhanced performance. We first validate the idea on a developed lightweight toy d etector which carries simplest notion of current state-of-the-art anchor based d etection models on challenging KITTI dataset, our method generates up to 15% boo st of mAP for the student model compared to the non-imitated counterpart. We the n extensively evaluate the method with Faster R-CNN model under various scenario s with common object detection benchmark of Pascal VOC and COCO, imitation allev iates up to 74% performance drop of student model compared to teacher. Codes rel eased at https://github.com/twangnh/Distilling-Object-Detectors
****

Centripetal SGD for Pruning Very Deep Convolutional Networks With Complicated St ructure

Xiaohan Ding, Guiguang Ding, Yuchen Guo, Jungong Han; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 494 3-4953

The redundancy is widely recognized in Convolutional Neural Networks (CNNs), whi ch enables to remove some unimportant filters from convolutional layers so as to slim the network with acceptable performance drop. Inspired by the linearity of convolution, we seek to make some filters increasingly close and eventually ide ntical for network slimming. To this end, we propose Centripetal SGD (C-SGD), a novel optimization method, which can train several filters to collapse into a si ngle point in the parameter hyperspace. When the training is completed, the remo val of the identical filters can trim the network with NO performance loss, thus no finetuning is needed. By doing so, we have partly solved an open problem of constrained filter pruning on CNNs with complicated structure, where some layers must be pruned following the others. Our experimental results on CIFAR-10 and I mageNet have justified the effectiveness of C-SGD-based filter pruning. Moreover , we have provided empirical evidences for the assumption that the redundancy in deep neural networks helps the convergence of training by showing that a redund ant CNN trained using C-SGD outperforms a normally trained counterpart with the equivalent width.
****

Knockoff Nets: Stealing Functionality of Black-Box Models

Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4954-49 63

Machine Learning (ML) models are increasingly deployed in the wild to perform a wide range of tasks. In this work, we ask to what extent can an adversary steal functionality of such "victim" models based solely on blackbox interactions: ima ge in, predictions out. In contrast to prior work, we study complex victim black box models, and an adversary lacking knowledge of train/test data used by the mo del, its internals, and semantics over model outputs. We formulate model functio nality stealing as a two-step approach: (i) querying a set of input images to th e blackbox model to obtain predictions; and (ii) training a "knockoff" with quer ied image-prediction pairs. We make multiple remarkable observations: (a) queryi ng random images from a different distribution than that of the blackbox trainin g data results in a well-performing knockoff; (b) this is possible even when the knockoff is represented using a different architecture; and (c) our reinforceme nt learning approach additionally improves query sample efficiency in certain se ttings and provides performance gains. We validate model functionality stealing on a range of datasets and tasks, as well as show that a reasonable knockoff of an image analysis API could be created for as little as 30.
****

Deep Embedding Learning With Discriminative Sampling Policy

Yueqi Duan, Lei Chen, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4964-4973
Deep embedding learning aims to learn a distance metric for effective similarity measurement, which has achieved promising performance in various tasks. As the vast majority of training samples produce gradients with magnitudes close to zero, hard example mining is usually employed to improve the effectiveness and efficiency of the training procedure. However, most existing sampling methods are designed by hand, which ignores the dependence between examples and suffer from exhaustive searching. In this paper, we propose a deep embedding with discriminative sampling policy (DE-DSP) learning framework by simultaneously training two models: a deep sampler network that learns effective sampling strategies, and a feature embedding that maps samples to the feature space. Rather than exhaustively calculating the hardness of all the examples for mining through forward-propagation, the deep sampler network exploits the strong prior of relations among samples to learn discriminative sampling policy in an more efficient manner. Experimental results demonstrate faster convergence and stronger discriminative power of our DE-DSP framework under different embedding objectives.
****

Hybrid Task Cascade for Instance Segmentation
Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4974-4983
Cascade is a classic yet powerful architecture that has boosted performance on various tasks. However, how to introduce cascade to instance segmentation remains an open question. A simple combination of Cascade R-CNN and Mask R-CNN only brings limited gain. In exploring a more effective approach, we find that the key to a successful instance segmentation cascade is to fully leverage the reciprocal relationship between detection and segmentation. In this work, we propose a new framework, Hybrid Task Cascade (HTC), which differs in two important aspects: (1) instead of performing cascaded refinement on these two tasks separately, it interweaves them for a joint multi-stage processing; (2) it adopts a fully convolutional branch to provide spatial context, which can help distinguishing hard foreground from cluttered background. Overall, this framework can learn more discriminative features progressively while integrating complementary features together in each stage. Without bells and whistles, a single HTC obtains 38.4% and 1.5% improvement over a strong Cascade Mask R-CNN baseline on MSCOCO dataset. Moreover, our overall system achieves 48.6 mask AP on the test-challenge split, ranking 1st in the COCO 2018 Challenge Object Detection Task. Code is available at https://github.com/open-mmlab/mmdetection.
****

Multi-Task Self-Supervised Object Detection via Recycling of Bounding Box Annotations
Wonhee Lee, Joonil Na, Gunhee Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4984-4993
In spite of recent enormous success of deep convolutional networks in object detection, they require a large amount of bounding box annotations, which are often time-consuming and error-prone to obtain. To make better use of given limited labels, we propose a novel object detection approach that takes advantage of both multi-task learning (MTL) and self-supervised learning (SSL). We propose a set of auxiliary tasks that help improve the accuracy of object detection. They create their own labels by recycling the bounding box labels (i.e. annotations of the main task) in an SSL manner, and are jointly trained with the object detection model in an MTL way. Our approach is integrable with any region proposal based detection models. We empirically validate that our approach effectively improves detection performance on various architectures and datasets. We test two state-of-the-art region proposal object detectors, including Faster R-CNN and R-FCN, with three CNN backbones of ResNet-101, Inception-ResNet-v2, and MobileNet on two benchmark datasets of PASCAL VOC and COCO.
****

ClusterNet: Deep Hierarchical Cluster Network With Rigorously Rotation-Invariant Representation for Point Cloud Analysis
Chao Chen, Guanbin Li, Ruijia Xu, Tianshui Chen, Meng Wang, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4994-5002
Current neural networks for 3D object recognition are vulnerable to 3D rotation. Existing works mostly rely on massive amounts of rotation-augmented data to alleviate the problem, which lacks solid guarantee of the 3D rotation invariance. In this paper, we address the issue by introducing a novel point cloud representation that can be mathematically proved rigorously rotation-invariant, i.e., identical point clouds in different orientations are unified as a unique and consistent representation. Moreover, the proposed representation is conditional information-lossless, because it retains all necessary information of point cloud except for orientation information. In addition, the proposed representation is complementary with existing network architectures for point cloud and fundamentally improves their robustness against rotation transformation. Finally, we propose a deep hierarchical cluster network called ClusterNet to better adapt to the proposed representation. We employ hierarchical clustering to explore and exploit the geometric structure of point cloud, which is embedded in a hierarchical structure tree. Extensive experimental results have shown that our proposed method greatly outperforms the state-of-the-arts in rotation robustness on rotation-augmented 3D object classification benchmarks.
****
Learning to Learn Relation for Important People Detection in Still Images
Wei-Hong Li, Fa-Ting Hong, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5003-5011
Humans can easily recognize the importance of people in social event images, and they always focus on the most important individuals. However, learning to learn the relation between people in an image, and inferring the most important person based on this relation, remains undeveloped. In this work, we propose a deep imPOrtance relatIon NeTwork (POINT) that combines both relation modeling and feature learning. In particular, we infer two types of interaction modules: the person-person interaction module that learns the interaction between people and the event-person interaction module that learns to describe how a person is involved in the event occurring in an image. We then estimate the importance relations among people from both interactions and encode the relation feature from the importance relations. In this way, POINT automatically learns several types of relation features in parallel, and we aggregate these relation features and the person's feature to form the importance feature for important people classification. Extensive experimental results show that our method is effective for important people detection and verify the efficacy of learning to learn relations for important people detection.
****
Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition
Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5012-5021
Learning subtle yet discriminative features (e.g., beak and eyes for a bird) plays a significant role in fine-grained image recognition. Existing attention-based approaches localize and amplify significant parts to learn fine-grained details, which often suffer from a limited number of parts and heavy computational cost. In this paper, we propose to learn such fine-grained features from hundreds of part proposals by Trilinear Attention Sampling Network (TASN) in an efficient teacher-student manner. Specifically, TASN consists of 1) a trilinear attention module, which generates attention maps by modeling the inter-channel relationships, 2) an attention-based sampler which highlights attended parts with high resolution, and 3) a feature distiller, which distills part features into an object-level feature by weight sharing and feature preserving strategies. Extensive experiments verify that TASN yields the best performance under the same settings wi

th the most competitive approaches, in iNaturalist-2017, CUB-Bird, and Stanford-Cars datasets.
****

Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning
Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, Matthew R. Scott; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5022-5030
A family of loss functions built on pair-based computation have been proposed in the literature which provide a myriad of solutions for deep metric learning. In this pa-per, we provide a general weighting framework for under-standing recent pair-based loss functions. Our contributions are three-fold: (1) we establish a General Pair Weighting (GPW) framework, which casts the sampling problem of deep metric learning into a unified view of pair weighting through gradient analysis, providing a powerful tool for understanding recent pair-based loss functions; (2) we show that with GPW, various existing pair-based methods can be compared and discussed comprehensively, with clear differences and key limitations identified; (3) we propose a new loss called multi-similarity loss (MS loss) under the GPW,which is implemented in two iterative steps (i.e., mining and weighting). This allows it to fully consider three similarities for pair weighting, providing a more principled approach for collecting and weighting informative pairs. Finally, the proposed MS loss obtains new state-of-the-art performance on four image retrieval benchmarks, where it outperforms the most recent approaches, such as ABE[14] and HTL[4], by a large margin, e.g.,60.6%-65.7%on CUB200,and 80.9%-88.0%on In-Shop Clothes Retrieval datasetat Recall@1.
****

Domain-Symmetric Networks for Adversarial Domain Adaptation
Yabin Zhang, Hui Tang, Kui Jia, Mingkui Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5031-5040
Unsupervised domain adaptation aims to learn a model of classifier for unlabeled samples on the target domain, given training data of labeled samples on the source domain. Impressive progress is made recently by learning invariant features via domain-adversarial training of deep networks. In spite of the recent progress, domain adaptation is still limited in achieving the invariance of feature distributions at a finer category level. To this end, we propose in this paper a new domain adaptation method called Domain-Symmetric Networks (SymNets). The proposed SymNet is based on a symmetric design of source and target task classifiers, based on which we also construct an additional classifier that shares with them its layer neurons. To train the SymNet, we propose a novel adversarial learning objective whose key design is based on a two-level domain confusion scheme, where the category-level confusion loss improves over the domain-level one by driving the learning of intermediate network features to be invariant at the corresponding categories of the two domains. Both domain discrimination and domain confusion are implemented based on the constructed additional classifier. Since target samples are unlabeled, we also propose a scheme of cross-domain training to help learn the target classifier. Careful ablation studies show the efficacy of our proposed method. In particular, based on commonly used base networks, our SymNets achieve the new state of the art on three benchmark domain adaptation datasets.
****

End-To-End Supervised Product Quantization for Image Search and Retrieval
Benjamin Klein, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5041-5050
Product Quantization, a dictionary based hashing method, is one of the leading unsupervised hashing techniques. While it ignores the labels, it harnesses the features to construct look up tables that can approximate the feature space. In recent years, several works have achieved state of the art results on hashing benchmarks by learning binary representations in a supervised manner. This work presents Deep Product Quantization (DPQ), a technique that leads to more accurate retrieval and classification than the latest state of the art methods, while having similar computational complexity and memory footprint as the Product Quantizat

ion method. To our knowledge, this is the first work to introduce a dictionary-based representation that is inspired by Product Quantization and which is learned end-to-end, and thus benefits from the supervised signal. DPQ explicitly learns soft and hard representations to enable an efficient and accurate asymmetric search, by using a straight-through estimator. Our method obtains state of the art results on an extensive array of retrieval and classification experiments.
****

Learning to Learn From Noisy Labeled Data
Junnan Li, Yongkang Wong, Qi Zhao, Mohan S. Kankanhalli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5051-5059
Despite the success of deep neural networks (DNNs) in image classification tasks, the human-level performance relies on massive training data with high-quality manual annotations, which are expensive and time-consuming to collect. There exist many inexpensive data sources on the web, but they tend to contain inaccurate labels. Training on noisy labeled datasets causes performance degradation because DNNs can easily overfit to the label noise. To overcome this problem, we propose a noise-tolerant training algorithm, where a meta-learning update is performed prior to conventional gradient update. The proposed meta-learning method simulates actual training by generating synthetic noisy labels, and train the model such that after one gradient update using each set of synthetic noisy labels, the model does not overfit to the specific noise. We conduct extensive experiments on the noisy CIFAR-10 dataset and the Clothing1M dataset. The results demonstrate the advantageous performance of the proposed method compared to several state-of-the-art baselines.
****

DSFD: Dual Shot Face Detector
Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, Feiyue Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5060-5069
Recently, Convolutional Neural Network (CNN) has achieved great success in face detection. However, it remains a challenging problem for the current face detection methods owing to high degree of variability in scale, pose, occlusion, expression, appearance and illumination. In this Paper, we propose a novel detection network named Dual Shot face Detector(DSFD). which inherits the architecture of SSD and introduces a Feature Enhance Module (FEM) for transferring the original feature maps to extend the single shot detector to dual shot detector. Specially, progressive anchor loss (PAL) computed by using two set of anchors is adopted to effectively facilitate the features. Additionally, we propose an improved anchor matching (IAM) method by integrating novel data augmentation techniques and anchor design strategy in our DSFD to provide better initialization for the regressor. Extensive experiments on popular benchmarks: WIDER FACE (easy: 0.966, medium: 0.957, hard: 0.904) and FDDB (discontinuous: 0.991, continuous: 0.862) demonstrate the superiority of DSFD over the state-of-the-art face detection methods (e.g., PyramidBox and SRN). Code will be made available upon publication.
****

Label Propagation for Deep Semi-Supervised Learning
Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Ondrej Chum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5070-5079
Semi-supervised learning is becoming increasingly important because it can combine data carefully labeled by humans with abundant unlabeled data to train deep neural networks. Classic methods on semi-supervised learning that have focused on transductive learning have not been fully exploited in the inductive framework followed by modern deep learning. The same holds for the manifold assumption---that similar examples should get the same prediction. In this work, we employ a transductive label propagation method that is based on the manifold assumption to make predictions on the entire dataset and use these predictions to generate pseudo-labels for the unlabeled data and train a deep neural network. At the core of the transductive method lies a nearest neighbor graph of the dataset that we

create based on the embeddings of the same network. Therefore our learning proce ss iterates between these two steps. We improve performance on several datasets especially in the few labels regime and show that our work is complementary to c urrent state of the art.
****

Deep Global Generalized Gaussian Networks
Qilong Wang, Peihua Li, Qinghua Hu, Pengfei Zhu, Wangmeng Zuo; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 5080-5088
Recently, global covariance pooling (GCP) has shown great advance in improving c lassification performance of deep convolutional neural networks (CNNs). However, existing deep GCP networks compute covariance pooling of convolutional activati ons with assumption that activations are sampled from Gaussian distributions, wh ich may not hold in practice and fails to fully characterize the statistics of a ctivations. To handle this issue, this paper proposes a novel deep global genera lized Gaussian network (3G-Net), whose core is to estimate a global covariance o f generalized Gaussian for modeling the last convolutional activations. Compared with GCP in Gaussian setting, our 3G-Net assumes the distribution of activation s follows a generalized Gaussian, which can capture more precise characteristics of activations. However, there exists no analytic solution for parameter estima tion of generalized Gaussian, making our 3G-Net challenging. To this end, we fir st present a novel regularized maximum likelihood estimator for robust estimatin g covariance of generalized Gaussian, which can be optimized by a modified itera tive re-weighted method. Then, to efficiently estimate the covariance of generai zed Gaussian under deep CNN architectures, we approximate this re-weighted metho d by developing an unrolling re-weighted module and a square root covariance lay er. In this way, 3GNet can be flexibly trained in an end-to-end manner. The expe riments are conducted on large-scale ImageNet-1K and Places365 datasets, and the results demonstrate our 3G-Net outperforms its counterparts while achieving ver y competitive performance to state-of-the-arts.
****

Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-Based Image Retr ieval
Anjan Dutta, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2019, pp. 5089-5098
Zero-shot sketch-based image retrieval (SBIR) is an emerging task in computer vi sion, allowing to retrieve natural images relevant to sketch queries that might not been seen in the training phase. Existing works either require aligned sketc h-image pairs or inefficient memory fusion layer for mapping the visual informat ion to a semantic space. In this work, we propose a semantically aligned paired cycle-consistent generative (SEM-PCYC) model for zero-shot SBIR, where each bran ch maps the visual information to a common semantic space via an adversarial tra ining. Each of these branches maintains a cycle consistency that only requires s upervision at category levels, and avoids the need of highly-priced aligned sket ch-image pairs. A classification criteria on the generators' outputs ensures the visual to semantic space mapping to be discriminating. Furthermore, we propose to combine textual and hierarchical side information via a feature selection aut o-encoder that selects discriminating side information within a same end-to-end model. Our results demonstrate a significant boost in zero-shot SBIR performance over the state-of-the-art on the challenging Sketchy and TU-Berlin datasets.
****

Context-Aware Crowd Counting
Weizhe Liu, Mathieu Salzmann, Pascal Fua; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5099-5108
State-of-the-art methods for counting people in crowded scenes rely on deep netw orks to estimate crowd density. They typically use the same filters over the who le image or over large image patches. Only then do they estimate local scale to compensate for perspective distortion. This is typically achieved by training an auxiliary classifier to select, for predefined image patches, the best kernel s ize among a limited set of choices. As such, these methods are not end-to-end tr

ainable and restricted in the scope of context they can leverage. In this paper, we introduce an end-to-end trainable deep architecture that combines features obtained using multiple receptive field sizes and learns the importance of each such feature at each image location. In other words, our approach adaptively encodes the scale of the contextual information required to accurately predict crowd density. This yields an algorithm that outperforms state-of-the-art crowd counting methods, especially when perspective effects are strong.
****

Detect-To-Retrieve: Efficient Regional Aggregation for Image Search
Marvin Teichmann, Andre Araujo, Menglong Zhu, Jack Sim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5109-5118
Retrieving object instances among cluttered scenes efficiently requires compact yet comprehensive regional image representations. Intuitively, object semantics can help build the index that focuses on the most relevant regions. However, due to the lack of bounding-box datasets for objects of interest among retrieval benchmarks, most recent work on regional representations has focused on either uniform or class-agnostic region selection. In this paper, we first fill the void by providing a new dataset of landmark bounding boxes, based on the Google Landmarks dataset, that includes 94k images with manually curated boxes from 15k unique landmarks. Then, we demonstrate how a trained landmark detector, using our new dataset, can be leveraged to index image regions and improve retrieval accuracy while being much more efficient than existing regional methods. In addition, we introduce a novel regional aggregated selective match kernel (R-ASMK) to effectively combine information from detected regions into an improved holistic image representation. R-ASMK boosts image retrieval accuracy substantially with no dimensionality increase, while even outperforming systems that index image regions independently. Our complete image retrieval system improves upon the previous state-of-the-art by significant margins on the Revisited Oxford and Paris datasets. Code and data will be released.
****

Towards Accurate One-Stage Object Detection With AP-Loss
Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, Junni Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5119-5127
One-stage object detectors are trained by optimizing classification-loss and localization-loss simultaneously, with the former suffering much from extreme foreground-background class imbalance issue due to the large number of anchors. This paper alleviates this issue by proposing a novel framework to replace the classification task in one-stage detectors with a ranking task, and adopting the Average-Precision loss (AP-loss) for the ranking problem. Due to its non-differentiability and non-convexity, the AP-loss cannot be optimized directly. For this purpose, we develop a novel optimization algorithm, which seamlessly combines the error-driven update scheme in perceptron learning and backpropagation algorithm in deep networks. We verify good convergence property of the proposed algorithm theoretically and empirically. Experimental results demonstrate notable performance improvement in state-of-the-art one-stage detectors based on AP-loss over different kinds of classification-losses on various benchmarks, without changing the network architectures.
****

On Exploring Undetermined Relationships for Visual Relationship Detection
Yibing Zhan, Jun Yu, Ting Yu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5128-5137
In visual relationship detection, human-notated relationships can be regarded as determinate relationships. However, there are still large amount of unlabeled data, such as object pairs with less significant relationships or even with no relationships. We refer to these unlabeled but potentially useful data as undetermined relationships. Although a vast body of literature exists, few methods exploit these undetermined relationships for visual relationship detection. In this paper, we explore the beneficial effect of undetermined relationships on visual

relationship detection. We propose a novel multi-modal feature based undetermined relationship learning network (MF-URLN) and achieve great improvements in relationship detection. In detail, our MF-URLN automatically generates undetermined relationships by comparing object pairs with human-notated data according to a designed criterion. Then, the MF-URLN extracts and fuses features of object pairs from three complementary modals: visual, spatial, and linguistic modals. Further, the MF-URLN proposes two correlated subnetworks: one subnetwork decides the determinate confidence, and the other predicts the relationships. We evaluate the MF-URLN on two datasets: the Visual Relationship Detection (VRD) and the Visual Genome (VG) datasets. The experimental results compared with state-of-the-art methods verify the significant improvements made by the undetermined relationships, e.g., the top-50 relation detection recall improves from 19.5% to 23.9% on the VRD dataset.

****

Learning Without Memorizing

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, Rama Chellappa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5138-5146

Incremental learning (IL) is an important task aimed at increasing the capability of a trained model, in terms of the number of classes recognizable by the model. The key problem in this task is the requirement of storing data (e.g. images) associated with existing classes, while teaching the classifier to learn new classes. However, this is impractical as it increases the memory requirement at every incremental step, which makes it impossible to implement IL algorithms on edge devices with limited memory. Hence, we propose a novel approach, called `Learning without Memorizing (LwM)', to preserve the information about existing (base) classes, without storing any of their data, while making the classifier progressively learn the new classes. In LwM, we present an information preserving penalty: Attention Distillation Loss (L_ AD), and demonstrate that penalizing the changes in classifiers' attention maps helps to retain information of the base classes, as new classes are added. We show that adding L_ AD to the distillation loss which is an existing information preserving loss consistently outperforms the state-of-the-art performance in the iILSVRC-small and iCIFAR-100 datasets in terms of the overall accuracy of base and incrementally learned classes.

****

Dynamic Recursive Neural Network

Qiushan Guo, Zhipeng Yu, Yichao Wu, Ding Liang, Haoyu Qin, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5147-5156

This paper proposes the dynamic recursive neural network (DRNN), which simplifies the duplicated building blocks in deep neural network. Different from forwarding through different blocks sequentially in previous networks, we demonstrate that the DRNN can achieve better performance with fewer blocks by employing block recursively. We further add a gate structure to each block, which can adaptively decide the loop times of recursive blocks to reduce the computational cost. Since the recursive networks are hard to train, we propose the Loopy Variable Batch Normalization (LVBN) to stabilize the volatile gradient. Further, we improve the LVBN to correct statistical bias caused by the gate structure. Experiments show that the DRNN reduces the parameters and computational cost and while outperforms the original model in term of the accuracy consistently on CIFAR-10 and ImageNet-1k. Lastly we visualize and discuss the relation between image saliency and the number of loop time.

****

Destruction and Construction Learning for Fine-Grained Image Recognition

Yue Chen, Yalong Bai, Wei Zhang, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5157-5166

Delicate feature representation about object parts plays a critical role in fine-grained recognition. For example, experts can even distinguish fine-grained objects relying only on object parts according to professional knowledge. In this paper, we propose a novel "Destruction and Construction Learning" (DCL) method to

enhance the difficulty of fine-grained recognition and exercise the classification model to acquire expert knowledge. Besides the standard classification backbone network, another "destruction and construction" stream is introduced to carefully "destruct" and then "reconstruct" the input image, for learning discriminative regions and features. More specifically, for "destruction", we first partition the input image into local regions and then shuffle them by a Region Confusion Mechanism (RCM). To correctly recognize these destructed images, the classification network has to pay more attention to discriminative regions for spotting the differences. To compensate the noises introduced by RCM, an adversarial loss, which distinguishes original images from destructed ones, is applied to reject noisy patterns introduced by RCM. For "construction", a region alignment network, which tries to restore the original spatial layout of local regions, is followed to model the semantic correlation among local regions. By jointly training with parameter sharing, our proposed DCL injects more discriminative local details to the classification network. Experimental results show that our proposed framework achieves state-of-the-art performance on three standard benchmarks. Moreover, our proposed method does not need any external knowledge during training, and there is no computation overhead at inference time except the standard classification network feed-forwarding. Source code: https://github.com/JDAI-CV/DCL.
****

Distraction-Aware Shadow Detection
Quanlong Zheng, Xiaotian Qiao, Ying Cao, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5167-5176
Shadow detection is an important and challenging task for scene understanding. Despite promising results from recent deep learning based methods. Existing works still struggle with ambiguous cases where the visual appearances of shadow and non-shadow regions are similar (referred to as distraction in our context). In this paper, we propose a Distraction-aware Shadow Detection Network (DSDNet) by explicitly learning and integrating the semantics of visual distraction regions in an end-to-end framework. At the core of our framework is a novel standalone, differentiable Distraction-aware Shadow (DS) module, which allows us to learn distraction-aware, discriminative features for robust shadow detection, by explicitly predicting false positives and false negatives. We conduct extensive experiments on three public shadow detection datasets, SBU, UCF and ISTD, to evaluate our method. Experimental results demonstrate that our model can boost shadow detection performance, by effectively suppressing the detection of false positives and false negatives, achieving state-of-the-art results.
****

Multi-Label Image Recognition With Graph Convolutional Networks
Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, Yanwen Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5177-5186
The task of multi-label image recognition is to predict a set of object labels that present in an image. As objects normally co-occur in an image, it is desirable to model the label dependencies to improve the recognition performance. To capture and explore such important dependencies, we propose a multi-label classification model based on Graph Convolutional Network (GCN). The model builds a directed graph over the object labels, where each node (label) is represented by word embeddings of a label, and GCN is learned to map this label graph into a set of inter-dependent object classifiers. These classifiers are applied to the image descriptors extracted by another sub-net, enabling the whole network to be end-to-end trainable. Furthermore, we propose a novel re-weighted scheme to create an effective label correlation matrix to guide information propagation among the nodes in GCN. Experiments on two multi-label image recognition datasets show that our approach obviously outperforms other existing state-of-the-art methods. In addition, visualization analyses reveal that the classifiers learned by our model maintain meaningful semantic topology.
****

High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detectio

n
Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, Yinan Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5187-5196
Object detection generally requires sliding-window classifiers in tradition or anchor-based predictions in modern deep learning approaches. However, either of these approaches requires tedious configurations in windows or anchors. In this paper, taking pedestrian detection as an example, we provide a new perspective where detecting objects is motivated as a high-level semantic feature detection task. Like edges, corners, blobs and other feature detectors, the proposed detector scans for feature points all over the image, for which the convolution is naturally suited. However, unlike these traditional low-level features, the proposed detector goes for a higher-level abstraction, that is, we are looking for central points where there are pedestrians, and modern deep models are already capable of such a high-level semantic abstraction. Besides, like blob detection, we also predict the scales of the pedestrian points, which is also a straightforward convolution. Therefore, in this paper, pedestrian detection is simplified as a straightforward center and scale prediction task through convolutions. This way, the proposed method enjoys an anchor-free setting. Though structurally simple, it presents competitive accuracy and good speed on challenging pedestrian detection benchmarks, and hence leading to a new attractive pedestrian detector. Code and models will be available at https://github.com/liuwei16/CSP.
****
RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection
Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, Alex M. Bronstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5197-5206
Distance metric learning (DML) has been successfully applied to object classification, both in the standard regime of rich training data and in the few-shot scenario, where each category is represented by only a few examples. In this work, we propose a new method for DML that simultaneously learns the backbone network parameters, the embedding space, and the multi-modal distribution of each of the training categories in that space, in a single end-to-end training process. Our approach outperforms state-of-the-art methods for DML-based object classification on a variety of standard fine-grained datasets. Furthermore, we demonstrate the effectiveness of our approach on the problem of few-shot object detection, by incorporating the proposed DML architecture as a classification head into a standard object detection model. We achieve the best results on the ImageNet-LOC dataset compared to strong baselines, when only a few training examples are available. We also offer the community a new episodic benchmark based on the ImageNet dataset for the few-shot object detection task.
****
Ranked List Loss for Deep Metric Learning
Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, Neil M. Robertson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5207-5216
The objective of deep metric learning (DML) is to learn embeddings that can capture semantic similarity information among data points. Existing pairwise or tripletwise loss functions used in DML are known to suffer from slow convergence due to a large proportion of trivial pairs or triplets as the model improves. To improve this, rankingmotivated structured losses are proposed recently to incorporate multiple examples and exploit the structured information among them. They converge faster and achieve state-of-the-art performance. In this work, we present two limitations of existing ranking-motivated structured losses and propose a novel ranked list loss to solve both of them. First, given a query, only a fraction of data points is incorporated to build the similarity structure. Consequently, some useful examples are ignored and the structure is less informative. To address this, we propose to build a setbased similarity structure by exploiting all instances in the gallery. The samples are split into a positive set and a nega

tive set. Our objective is to make the query closer to the positive set than to the negative set by a margin. Second, previous methods aim to pull positive pairs as close as possible in the embedding space. As a result, the intraclass data distribution might be dropped. In contrast, we propose to learn a hypersphere for each class in order to preserve the similarity structure inside it. Our extensive experiments show that the proposed method achieves state-of-the-art performance on three widely used benchmarks.

****

CANet: Class-Agnostic Segmentation Networks With Iterative Refinement and Attentive Few-Shot Learning

Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5217-5226

Recent progress in semantic segmentation is driven by deep Convolutional Neural Networks and large-scale labeled image datasets. However, data labeling for pixel-wise segmentation is tedious and costly. Moreover, a trained model can only make predictions within a set of pre-defined classes. In this paper, we present CANet, a class-agnostic segmentation network that performs few-shot segmentation on new classes with only a few annotated images available. Our network consists of a two-branch dense comparison module which performs multi-level feature comparison between the support image and the query image, and an iterative optimization module which iteratively refines the predicted results. Furthermore, we introduce an attention mechanism to effectively fuse information from multiple support examples under the setting of k-shot learning. Experiments on PASCAL VOC 2012 show that our method achieves a mean Intersection-over-Union score of 55.4% for 1-shot segmentation and 57.1% for 5-shot segmentation, outperforming state-of-the-art methods by a large margin of 14.6% and 13.2%, respectively.

****

Precise Detection in Densely Packed Scenes

Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, Tal Hassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5227-5236

Man-made scenes are often densely packed, containing numerous objects, often identical, positioned in close proximity. We show that precise object detection in such scenes remains a challenging frontier even for state-of-the-art object detectors. We propose a novel, deep-learning based method for precise object detection, designed for such challenging settings. Our contributions include: (1) A layer for estimating the Jaccard index as a detection quality score; (2) a novel EM merging unit, which uses our quality scores to resolve detection overlap ambiguities; finally, (3) an extensive, annotated data set, SKU-110K, representing packed retail environments, released for training and testing under such extreme settings. Detection tests on SKU-110K, and counting tests on the CARPK and PUCPR+, show our method to outperform existing state-of-the-art with substantial margins.

****

KE-GAN: Knowledge Embedded Generative Adversarial Networks for Semi-Supervised Scene Parsing

Mengshi Qi, Yunhong Wang, Jie Qin, Annan Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5237-5246

In recent years, scene parsing has captured increasing attention in computer vision. Previous works have demonstrated promising performance in this task. However, they mainly utilize holistic features, whilst neglecting the rich semantic knowledge and inter-object relationships in the scene. In addition, these methods usually require a large number of pixel-level annotations, which is too expensive in practice. In this paper, we propose a novel Knowledge Embedded Generative Adversarial Networks, dubbed as KE-GAN, to tackle the challenging problem in a semi-supervised fashion. KE-GAN captures semantic consistencies of different categories by devising a Knowledge Graph from the large-scale text corpus. In addition to readily-available unlabeled data, we generate synthetic images to unveil rich structural information underlying the images. Moreover, a pyramid architectur

e is incorporated into the discriminator to acquire multi-scale contextual information for better parsing results. Extensive experimental results on four standard benchmarks demonstrate that KE-GAN is capable of improving semantic consistencies and learning better representations for scene parsing, resulting in the state-of-the-art performance.
****

Fast User-Guided Video Object Segmentation by Interaction-And-Propagation Networks
Seoung Wug Oh, Joon-Young Lee, Ning Xu, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5247-5256
We present a deep learning method for the interactive video object segmentation. Our method is built upon two core operations, interaction and propagation, and each operation is conducted by Convolutional Neural Networks. The two networks are connected both internally and externally so that the networks are trained jointly and interact with each other to solve the complex video object segmentation problem. We propose a new multi-round training scheme for the interactive video object segmentation so that the networks can learn how to understand the user's intention and update incorrect estimations during the training. At the testing time, our method produces high-quality results and also runs fast enough to work with users interactively. We evaluated the proposed method quantitatively on the interactive track benchmark at the DAVIS Challenge 2018. We outperformed other competing methods by a significant margin in both the speed and the accuracy. We also demonstrated that our method works well with real user interactions.
****

Fast Interactive Object Annotation With Curve-GCN
Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5257-5266
Manually labeling objects by tracing their boundaries is a laborious process. In Polygon-RNN++, the authors proposed Polygon-RNN that produces polygonal annotations in a recurrent manner using a CNN-RNN architecture, allowing interactive correction via humans-in-the-loop. We propose a new framework that alleviates the sequential nature of Polygon-RNN, by predicting all vertices simultaneously using a Graph Convolutional Network (GCN). Our model is trained end-to-end, and runs in real time. It supports object annotation by either polygons or splines, facilitating labeling efficiency for both line-based and curved objects. We show that Curve-GCN outperforms all existing approaches in automatic mode, including the powerful DeepLab, and is significantly more efficient in interactive mode than Polygon-RNN++. Our model runs at 29.3ms in automatic, and 2.6ms in interactive mode, making it 10x and 100x faster than Polygon-RNN++.
****

FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference
Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, Sungroh Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5267-5276
The main obstacle to weakly supervised semantic image segmentation is the difficulty of obtaining pixel-level information from coarse image-level annotations. Most methods based on image-level annotations use localization maps obtained from the classifier, but these only focus on the small discriminative parts of objects and do not capture precise boundaries. FickleNet explores diverse combinations of locations on feature maps created by generic deep neural networks. It selects hidden units randomly and then uses them to obtain activation scores for image classification. FickleNet implicitly learns the coherence of each location in the feature maps, resulting in a localization map which identifies both discriminative and other parts of objects. The ensemble effects are obtained from a single network by selecting random hidden unit pairs, which means that a variety of localization maps are generated from a single image. Our approach does not require any additional training steps and only adds a simple layer to a standard conv

olutional neural network; nevertheless it outperforms recent comparable techniqu
es on the Pascal VOC 2012 benchmark in both weakly and semi-supervised settings.
****

RVOS: End-To-End Recurrent Network for Video Object Segmentation
Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marque
s, Xavier Giro-i-Nieto; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2019, pp. 5277-5286
Multiple object video object segmentation is a challenging task, specially for t
he zero-shot case, when no object mask is given at the initial frame and the mod
el has to find the objects to be segmented along the sequence. In our work, we p
ropose a Recurrent network for multiple object Video Object Segmentation (RVOS)
that is fully end-to-end trainable. Our model incorporates recurrence on two dif
ferent domains: (i) the spatial, which allows to discover the different object i
nstances within a frame, and (ii) the temporal, which allows to keep the coheren
ce of the segmented objects along time. We train RVOS for zero-shot video object
segmentation and are the first ones to report quantitative results for DAVIS-20
17 and YouTube-VOS benchmarks. Further, we adapt RVOS for one-shot video object
segmentation by using the masks obtained in previous time steps as inputs to be
processed by the recurrent module. Our model reaches comparable results to state
-of-the-art techniques in YouTube-VOS benchmark and outperforms all previous vid
eo object segmentation methods not using online learning in the DAVIS-2017 bench
mark. Moreover, our model achieves faster inference runtimes than previous metho
ds, reaching 44ms/frame on a P100 GPU.
****

DeepFlux for Skeletons in the Wild
Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, Sven Dickinson, Kalee
m Siddiqi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2019, pp. 5287-5296
Computing object skeletons in natural images is challenging, owing to large vari
ations in object appearance and scale, and the complexity of handling background
clutter. Many recent methods frame object skeleton detection as a binary pixel
classification problem, which is similar in spirit to learning-based edge detect
ion, as well as to semantic segmentation methods. In the present article, we dep
art from this strategy by training a CNN to predict a two-dimensional vector fie
ld, which maps each scene point to a candidate skeleton pixel, in the spirit of
flux-based skeletonization algorithms. This "image context flux" representation
has two major advantages over previous approaches. First, it explicitly encodes
the relative position of skeletal pixels to semantically meaningful entities, su
ch as the image points in their spatial context, and hence also the implied obje
ct boundaries. Second, since the skeleton detection context is a region-based ve
ctor field, it is better able to cope with object parts of large width. We evalu
ate the proposed method on three benchmark datasets for skeleton detection and t
wo for symmetry detection, achieving consistently superior performance over stat
e-of-the-art methods.
****

Interactive Image Segmentation via Backpropagating Refinement Scheme
Won-Dong Jang, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2019, pp. 5297-5306
An interactive image segmentation algorithm, which accepts user-annotations abou
t a target object and the background, is proposed in this work. We convert user-
annotations into interaction maps by measuring distances of each pixel to the an
notated locations. Then, we perform the forward pass in a convolutional neural n
etwork, which outputs an initial segmentation map. However, the user-annotated l
ocations can be mislabeled in the initial result. Therefore, we develop the back
propagating refinement scheme (BRS), which corrects the mislabeled pixels. Exper
imental results demonstrate that the proposed algorithm outperforms the conventi
onal algorithms on four challenging datasets. Furthermore, we demonstrate the ge
nerality and applicability of BRS in other computer vision tasks, by transformin
g existing convolutional neural networks into user-interactive ones.
****

Scene Parsing via Integrated Classification Model and Variance-Based Regularizat
ion
Hengcan Shi, Hongliang Li, Qingbo Wu, Zichen Song; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5307-5
316
Scene Parsing is a challenging task in computer vision, which can be formulated
as a pixel-wise classification problem. Existing deep-learning-based methods usu
ally use one general classifier to recognize all object categories. However, the
 general classifier easily makes some mistakes in dealing with some confusing ca
tegories that share similar appearances or semantics. In this paper, we propose
an integrated classification model and a variance-based regularization to achiev
e more accurate classifications. On the one hand, the integrated classification
model contains multiple classifiers, not only the general classifier but also a
refinement classifier to distinguish the confusing categories. On the other hand
, the variance-based regularization differentiates the scores of all categories
as large as possible to reduce misclassifications. Specifically, the integrated
classification model includes three steps. The first is to extract the features
of each pixel. Based on the features, the second step is to classify each pixel
across all categories to generate a preliminary classification result. In the th
ird step, we leverage a refinement classifier to refine the classification resul
t, focusing on differentiating the high-preliminary-score categories. An integra
ted loss with the variance-based regularization is used to train the model. Exte
nsive experiments on three common scene parsing datasets demonstrate the effecti
veness of the proposed method.
****

RAVEN: A Dataset for Relational and Analogical Visual REasoNing
Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, Song-Chun Zhu; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019,
 pp. 5317-5327
Dramatic progress has been witnessed in basic vision tasks involving low-level p
erception, such as object recognition, detection, and tracking. Unfortunately, t
here is still enormous performance gap between artificial vision systems and hum
an intelligence in terms of higher-level vision problems, especially ones involv
ing reasoning. Earlier attempts in equipping machines with high-level reasoning
have hovered around Visual Question Answering (VQA), one typical task associatin
g vision and language understanding. In this work, we propose a new dataset, bui
lt in the context of Raven's Progressive Matrices (RPM) and aimed at lifting mac
hine intelligence by associating vision with structural, relational, and analogi
cal reasoning in a hierarchical representation. Unlike previous works in measuri
ng abstract reasoning using RPM, we establish a semantic link between vision and
 reasoning by providing structure representation. This addition enables a new ty
pe of abstract reasoning by jointly operating on the structure representation. M
achine reasoning ability using modern computer vision is evaluated in this newly
 proposed dataset. Additionally, we also provide human performance as a referenc
e. Finally, we show consistent improvement across all models by incorporating a
simple neural module that combines visual understanding and structure reasoning.
****

Surface Reconstruction From Normals: A Robust DGP-Based Discontinuity Preservati
on Approach
Wuyuan Xie, Miaohui Wang, Mingqiang Wei, Jianmin Jiang, Jing Qin; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2019, pp. 5328-5336
In 3D surface reconstruction from normals, discontinuity preservation is an impo
rtant but challenging task. However, existing studies fail to address the discon
tinuous normal maps by enforcing the surface integrability in the continuous dom
ain. This paper introduces a robust approach to preserve the surface discontinui
ty in the discrete geometry way. Firstly, we design two representative normal in
compatibility features and propose an efficient discontinuity detection scheme t
o determine the splitting pattern for a discrete mesh. Secondly, we model the di
scontinuity preservation problem as a light-weight energy optimization framework

by jointly considering the discontinuity detection and the overall reconstruction error. Lastly, we further shrink the feasible solution space to reduce the complexity based on the prior knowledge. Experiments show that the proposed method achieves the best performance on an extensive 3D dataset compared with the state-of-the-arts in terms of mean angular error and computational complexity.
****

DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images
Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, Ping Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5337-5345
Understanding fashion images has been advanced by benchmarks with rich annotations such as DeepFashion, whose labels include clothing categories, landmarks, and consumer-commercial image pairs. However, DeepFashion has nonnegligible issues such as single clothing-item per image, sparse landmarks (4 8 only), and no per-pixel masks, making it had significant gap from real-world scenarios. We fill in the gap by presenting DeepFashion2 to address these issues. It is a versatile benchmark of four tasks including clothes detection, pose estimation, segmentation, and retrieval. It has 801K clothing items where each item has rich annotations such as style, scale, view- point, occlusion, bounding box, dense landmarks (e.g. 39 for 'long sleeve outwear' and 15 for 'vest'), and masks. There are also 873K Commercial-Consumer clothes pairs. The annotations of DeepFashion2 are much larger than its counterparts such as 8x of FashionAI Global Challenge. A strong baseline is proposed, called Match R- CNN, which builds upon Mask R-CNN to solve the above four tasks in an end-to-end manner. Extensive evaluations are conducted with different criterions in Deep- Fashion2. DeepFashion2 Dataset will be released at : https://github.com/switchablenorms/DeepFashion2
****

Jumping Manifolds: Geometry Aware Dense Non-Rigid Structure From Motion
Suryansh Kumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5346-5355
Given dense image feature correspondences of a non-rigidly moving object across multiple frames, this paper proposes an algorithm to estimate its 3D shape for each frame. To solve this problem accurately, the recent state-of-the-art algorithm reduces this task to set of local linear subspace reconstruction and clustering problem using Grassmann manifold representation [34]. Unfortunately, their method missed on some of the critical issues associated with the modeling of surface deformations, for e.g., the dependence of a local surface deformation on its neighbors. Furthermore, their representation to group high dimensional data points inevitably introduce the drawbacks of categorizing samples on the high-dimensional Grassmann manifold [32, 31]. Hence, to deal with such limitations with [34], we propose an algorithm that jointly exploits the benefit of high-dimensional Grassmann manifold to perform reconstruction, and its equivalent lower-dimensional representation to infer suitable clusters. To accomplish this, we project each Grassmannians onto a lower-dimensional Grassmann manifold which preserves and respects the deformation of the structure w.r.t its neighbors. These Grassmann points in the lower-dimension then act as a representative for the selection of high-dimensional Grassmann samples to perform each local reconstruction. In practice, our algorithm provides a geometrically efficient way to solve dense NRSfM by switching between manifolds based on its benefit and usage. Experimental results show that the proposed algorithm is very effective in handling noise with reconstruction accuracy as good as or better than the competing methods.
****

LVIS: A Dataset for Large Vocabulary Instance Segmentation
Agrim Gupta, Piotr Dollar, Ross Girshick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5356-5364
Progress on object detection is enabled by datasets that focus the research community's attention on open challenges. This process led us from simple images to complex scenes and from bounding boxes to segmentation masks. In this work, we introduce LVIS (pronounced 'el-vis'): a new dataset for Large Vocabulary Instance

Segmentation. We plan to collect 2.2 million high-quality instance segmentation masks for over 1000 entry-level object categories in 164k images. Due to the Zipfian distribution of categories in natural images, LVIS naturally has a long tail of categories with few training samples. Given that state-of-the-art deep learning methods for object detection perform poorly in the low-sample regime, we believe that our dataset poses an important and exciting new scientific challenge. LVIS is available at http://www.lvisdataset.org.

****

Fast Object Class Labelling via Speech

Michael Gygli, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5365-5373

Object class labelling is the task of annotating images with labels on the presence or absence of objects from a given class vocabulary. Simply asking one yes-no question per class, however, has a cost that is linear in the vocabulary size and is thus inefficient for large vocabularies. Modern approaches rely on a hierarchical organization of the vocabulary to reduce annotation time, but remain expensive (several minutes per image for the 200 classes in ILSVRC). Instead, we propose a new interface where classes are annotated via speech. Speaking is fast and allows for direct access to the class name, without searching through a list or hierarchy. As additional advantages, annotators can simultaneously speak and scan the image for objects, the interface can be kept extremely simple, and using it requires less mouse movement. As annotators using our interface should only say words from a given class vocabulary, we propose a dedicated task to train them to do so. Through experiments on COCO and ILSVRC, we show our method yields high-quality annotations at 2.3x -14.9x less annotation time than existing methods.

****

LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5374-5383

In this paper, we present LaSOT, a high-quality benchmark for Large-scale Single Object Tracking. LaSOT consists of 1,400 sequences with more than 3.5M frames in total. Each frame in these sequences is carefully and manually annotated with a bounding box, making LaSOT the largest, to the best of our knowledge, densely annotated tracking benchmark. The average video length of LaSOT is more than 2,500 frames, and each sequence comprises various challenges deriving from the wild where target objects may disappear and re-appear again in the view. By releasing LaSOT, we expect to provide the community with a large-scale dedicated benchmark with high quality for both the training of deep trackers and the veritable evaluation of tracking algorithms. Moreover, considering the close connections of visual appearance and natural language, we enrich LaSOT by providing additional language specification, aiming at encouraging the exploration of natural linguistic feature for tracking. A thorough experimental evaluation of 35 tracking algorithms on LaSOT is presented with detailed analysis, and the results demonstrate that there is still a big room for improvements.

****

Creative Flow+ Dataset

Maria Shugrina, Ziheng Liang, Amlan Kar, Jiaman Li, Angad Singh, Karan Singh, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5384-5393

We present the Creative Flow+ Dataset, the first diverse multi-style artistic video dataset richly labeled with per-pixel optical flow, occlusions, correspondences, segmentation labels, normals, and depth. Our dataset includes 3000 animated sequences rendered using styles randomly selected from 40 textured line styles and 38 shading styles, spanning the range between flat cartoon fill and wildly sketchy shading. Our dataset includes 124K+ train set frames and 10K test set frames rendered at 1500x1500 resolution, far surpassing the largest available optical flow datasets in size. While modern techniques for tasks such as optical flow estimation achieve impressive performance on realistic images and video, today

there is no way to gauge their performance on non-photorealistic images. Creativ
e Flow+ poses a new challenge to generalize real-world Computer Vision to messy
stylized content. We show that learning-based optical flow methods fail to gener
alize to this data and struggle to compete with classical approaches, and invite
 new research in this area. Our dataset and a new optical flow benchmark will b
e publicly available at: www.cs.toronto.edu/creativeflow/. We further release th
e complete dataset creation pipeline, allowing the community to generate and sty
lize their own data on demand.
****

Weakly Supervised Open-Set Domain Adaptation by Dual-Domain Collaboration
Shuhan Tan, Jiening Jiao, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5394-5403
In conventional domain adaptation, a critical assumption is that there exists a
fully labeled domain (source) that contains the same label space as another unla
beled or scarcely labeled domain (target). However, in the real world, there oft
en exist application scenarios in which both domains are partially labeled and n
ot all classes are shared between these two domains. Thus, it is meaningful to l
et partially labeled domains learn from each other to classify all the unlabeled
 samples in each domain under an open-set setting. We consider this problem as w
eakly supervised open-set domain adaptation. To address this practical setting,
we propose the Collaborative Distribution Alignment (CDA) method, which performs
 knowledge transfer bilaterally and works collaboratively to classify unlabeled
data and identify outlier samples. Extensive experiments on the Office benchmark
 and an application on person reidentification show that our method achieves sta
te-of-the-art performance.
****

A Neurobiological Evaluation Metric for Neural Network Model Search
Nathaniel Blanchard, Jeffery Kinnison, Brandon RichardWebster, Pouya Bashivan
, Walter J. Scheirer; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2019, pp. 5404-5413
Neuroscience theory posits that the brain's visual system coarsely identifies br
oad object categories via neural activation patterns, with similar objects produ
cing similar neural responses. Artificial neural networks also have internal act
ivation behavior in response to stimuli. We hypothesize that networks exhibiting
 brain-like activation behavior will demonstrate brain-like characteristics, e.g
., stronger generalization capabilities. In this paper we introduce a human-mode
l similarity (HMS) metric, which quantifies the similarity of human fMRI and net
work activation behavior. To calculate HMS, representational dissimilarity matri
ces (RDMs) are created as abstractions of activation behavior, measured by the c
orrelations of activations to stimulus pairs. HMS is then the correlation betwee
n the fMRI RDM and the neural network RDM across all stimulus pairs. We test the
 metric on unsupervised predictive coding networks, which specifically model vis
ual perception, and assess the metric for statistical significance over a large
range of hyperparameters. Our experiments show that networks with increased huma
n-model similarity are correlated with better performance on two computer vision
 tasks: next frame prediction and object matching accuracy. Further, HMS identif
ies networks with high performance on both tasks. An unexpected secondary findin
g is that the metric can be employed during training as an early-stopping mechan
ism.
****

Iterative Projection and Matching: Finding Structure-Preserving Representatives
and Its Application to Computer Vision
Alireza Zaeemzadeh, Mohsen Joneidi, Nazanin Rahnavard, Mubarak Shah; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2019, pp. 5414-5423
The goal of data selection is to capture the most structural information from a
set of data. This paper presents a fast and accurate data selection method, in w
hich the selected samples are optimized to span the subspace of all data. We pro
pose a new selection algorithm, referred to as iterative projection and matching
 (IPM), with linear complexity w.r.t. the number of data, and without any parame

ter to be tuned. In our algorithm, at each iteration, the maximum information from the structure of the data is captured by one selected sample, and the captured information is neglected in the next iterations by projection on the null-space of previously selected samples. The computational efficiency and the selection accuracy of our proposed algorithm outperform those of the conventional methods. Furthermore, the superiority of the proposed algorithm is shown on active learning for video action recognition dataset on UCF-101; learning using representatives on ImageNet; training a generative adversarial network (GAN) to generate multi-view images from a single-view input on CMU Multi-PIE dataset; and video summarization on UTE Egocentric dataset.

****

Efficient Multi-Domain Learning by Covariance Normalization

Yunsheng Li, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5424-5433

The problem of multi-domain learning of deep networks is considered. An adaptive layer is induced per target domain and a novel procedure, denoted covariance normalization (CovNorm), proposed to reduce its parameters. CovNorm is a data driven method of fairly simple implementation, requiring two principal component analyzes (PCA) and fine-tuning of a mini-adaptation layer. Nevertheless, it is shown, both theoretically and experimentally, to have several advantages over previous approaches, such as batch normalization or geometric matrix approximations. Furthermore, CovNorm can be deployed both when target datasets are available sequentially or simultaneously. Experiments show that, in both cases, it has performance comparable to a fully fine-tuned network, using as few as 0.13% of the corresponding parameters per target domain.

****

Predicting Visible Image Differences Under Varying Display Brightness and Viewing Distance

Nanyang Ye, Krzysztof Wolski, Rafal K. Mantiuk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5434-5442

Numerous applications require a robust metric that can predict whether image differences are visible or not. However, the accuracy of existing white-box visibility metrics, such as HDR-VDP, is often not good enough. CNN-based black-box visibility metrics have proven to be more accurate, but they cannot account for differences in viewing conditions, such as display brightness and viewing distance. In this paper, we propose a CNN-based visibility metric, which maintains the accuracy of deep network solutions and accounts for viewing conditions. To achieve this, we extend the existing dataset of locally visible differences (LocVis) with a new set of measurements, collected considering aforementioned viewing conditions. Then, we develop a hybrid model that combines white-box processing stages for modeling the effects of luminance masking and contrast sensitivity, with a black-box deep neural network. We demonstrate that the novel hybrid model can handle the change of viewing conditions correctly and outperforms state-of-the-art metrics.

****

A Bayesian Perspective on the Deep Image Prior

Zezhou Cheng, Matheus Gadelha, Subhransu Maji, Daniel Sheldon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5443-5451

The deep image prior was recently introduced as a prior for natural images. It represents images as the output of a convolutional network with random inputs. For "inference", gradient descent is performed to adjust network parameters to make the output match observations. This approach yields good performance on a range of image reconstruction tasks. We show that the deep image prior is asymptotically equivalent to a stationary Gaussian process prior in the limit as the number of channels in each layer of the network goes to infinity, and derive the corresponding kernel. This informs a Bayesian approach to inference. We show that by conducting posterior inference using stochastic gradient Langevin dynamics we avoid the need for early stopping, which is a drawback of the current approach, and improve results for denoising and impainting tasks. We illustrate these intui

tions on a number of 1D and 2D signal reconstruction tasks.
****

ApolloCar3D: A Large 3D Car Instance Understanding Benchmark for Autonomous Driving
Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, Ruigang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5452-5462
Autonomous driving has attracted remarkable attention from both industry and academia. An important task is to estimate 3D properties (e.g. translation, rotation and shape) of a moving or parked vehicle on the road. This task, while critical, is still under-researched in the computer vision community - partially owing to the lack of large scale and fully-annotated 3D car database suitable for autonomous driving research. In this paper, we contribute the first large scale database suitable for 3D car instance understanding - ApolloCar3D. The dataset contains 5,277 driving images and over 60K car instances, where each car is fitted with an industry-grade 3D CAD model with absolute model size and semantically labelled keypoints. This dataset is above 20x larger than PASCAL3D+ and KITTI, the current state-of-the-art. To enable efficient labelling in 3D, we build a pipeline by considering 2D-3D keypoint correspondences for a single instance and 3D relationship among multiple instances. Equipped with such dataset, we build various baseline algorithms with the state-of-the-art deep convolutional neural networks. Specifically, we first segment each car with a pre-trained Mask R-CNN, and then regress towards its 3D pose and shape based on a deformable 3D car model with or without using semantic keypoints. We show that using keypoints significantly improves fitting performance. Finally, we develop a new 3D metric jointly considering 3D pose and 3D shape, allowing for comprehensive evaluation and ablation study.
****

Compressing Unknown Images With Product Quantizer for Efficient Zero-Shot Classification
Jin Li, Xuguang Lan, Yang Liu, Le Wang, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5463-5472
For Zero-Shot Learning (ZSL), the Nearest Neighbor (NN) search is generally conducted for classification, which may cause unacceptable computational complexity for large-scale datasets. To compress zero-shot classes by the trained quantizer for efficient search, it tends to induce large quantization error because distributions between seen and unseen classes are different. However, as semantic attributes of classes are available in ZSL, both seen and unseen classes have the same distribution for one specific property, e.g., animals have or not have spots. Based on this intuition, a Product Quantization Zero-Shot Learning (PQZSL) method is proposed to learn embeddings as well as quantizers to compress visual features into compact codes for Approximate NN (ANN) search. Particularly, visual features are projected into an orthogonal semantic space, and then the Product Quantization (PQ) is utilized to quantize individual properties. Experimental results on five benchmark datasets demonstrate that unseen classes are represented by the Cartesian product of quantized properties with little quantization error. As classes in orthogonal common space are more discriminative, the classification based on PQZSL achieves state-of-the-art performance in Generalized Zero-Shot Learning (GZSL) task, meanwhile, the speed of ANN search is 10-100 times higher than traditional NN search.
****

Self-Supervised Convolutional Subspace Clustering Network
Junjian Zhang, Chun-Guang Li, Chong You, Xianbiao Qi, Honggang Zhang, Jun Guo, Zhouchen Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5473-5482
Subspace clustering methods based on data self-expression have become very popular for learning from data that lie in a union of low-dimensional linear subspaces. However, the applicability of subspace clustering has been limited because practical visual data in raw form do not necessarily lie in such linear subspaces.

On the other hand, while Convolutional Neural Network (ConvNet) has been demonstrated to be a powerful tool for extracting discriminative features from visual data, training such a ConvNet usually requires a large amount of labeled data, which are unavailable in subspace clustering applications. To achieve simultaneous feature learning and subspace clustering, we propose an end-to-end trainable framework, called Self-Supervised Convolutional Subspace Clustering Network (S^2ConvSCN), that combines a ConvNet module (for feature learning), a self-expression module (for subspace clustering) and a spectral clustering module (for self-supervision) into a joint optimization framework. Particularly, we introduce a dual self-supervision that exploits the output of spectral clustering to supervise the training of the feature learning module (via a classification loss) and the self-expression module (via a spectral clustering loss). Our experiments on four benchmark datasets show the effectiveness of the dual self-supervision and demonstrate superior performance of our proposed approach.

****

Multi-Scale Geometric Consistency Guided Multi-View Stereo

Qingshan Xu, Wenbing Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5483-5492

In this paper, we propose an efficient multi-scale geometric consistency guided multi-view stereo method for accurate and complete depth map estimation. We first present our basic multi-view stereo method with Adaptive Checkerboard sampling and Multi-Hypothesis joint view selection (ACMH). It leverages structured region information to sample better candidate hypotheses for propagation and infer the aggregation view subset at each pixel. For the depth estimation of low-textured areas, we further propose to combine ACMH with multi-scale geometric consistency guidance (ACMM) to obtain the reliable depth estimates for low-textured areas at coarser scales and guarantee that they can be propagated to finer scales. To correct the erroneous estimates propagated from the coarser scales, we present a novel detail restorer. Experiments on extensive datasets show our method achieves state-of-the-art performance, recovering the depth estimation not only in low-textured areas but also in details.

****

Privacy Preserving Image-Based Localization

Pablo Speciale, Johannes L. Schonberger, Sing Bing Kang, Sudipta N. Sinha, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5493-5503

Image-based localization is a core component of many augmented/mixed reality (AR/MR) and autonomous robotic systems. Current localization systems rely on the persistent storage of 3D point clouds of the scene to enable camera pose estimation, but such data reveals potentially sensitive scene information. This gives rise to significant privacy risks, especially as for many applications 3D mapping is a background process that the user might not be fully aware of. We pose the following question: How can we avoid disclosing confidential information about the captured 3D scene, and yet allow reliable camera pose estimation? This paper proposes the first solution to what we call privacy preserving image-based localization. The key idea of our approach is to lift the map representation from a 3D point cloud to a 3D line cloud. This novel representation obfuscates the underlying scene geometry while providing sufficient geometric constraints to enable robust and accurate 6-DOF camera pose estimation. Extensive experiments on several datasets and localization scenarios underline the high practical relevance of our proposed approach.

****

SimulCap : Single-View Human Performance Capture With Cloth Simulation

Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5504-5514

This paper proposes a new method for live free-viewpoint human performance capture with dynamic details (e.g., cloth wrinkles) using a single RGBD camera. Our main contributions are: (i) a multi-layer representation of garments and body, and (ii) a physics-based performance capture procedure. We first digitize the perf

ormer using multi-layer surface representation, which includes the undressed bod y surface and separate clothing meshes. For performance capture, we perform skel eton tracking, cloth simulation, and iterative depth fitting sequentially for th e incoming frame. By incorporating cloth simulation into the performance capture pipeline, we can simulate plausible cloth dynamics and cloth-body interactions even in the occluded regions, which was not possible in previous capture methods . Moreover, by formulating depth fitting as a physical process, our system produ ces cloth tracking results consistent with the depth observation while still mai ntaining physical constraints. Results and evaluations show the effectiveness of our method. Our method also enables new types of applications such as cloth ret argeting, free-viewpoint video rendering and animations.
****

Hierarchical Deep Stereo Matching on High-Resolution Images
Gengshan Yang, Joshua Manela, Michael Happold, Deva Ramanan; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5515-5524
We explore the problem of real-time stereo matching on high-res imagery. Many st ate-of-the-art (SOTA) methods struggle to process high-res imagery because of me mory constraints or speed limitations. To address this issue, we propose an end- to-end framework that searches for correspondences incrementally over a coarse-t o-fine hierarchy. Because high-res stereo datasets are relatively rare, we intro duce a dataset with high-res stereo pairs for both training and evaluation. Our approach achieved SOTA performance on Middlebury-v3 and KITTI-15 while running s ignificantly faster than its competitors. The hierarchical design also naturally allows for anytime on-demand reports of disparity by capping intermediate coars e results, allowing us to accurately predict disparity for near-range structures with low latency (30ms). We demonstrate that the performance-vs-speed tradeoff afforded by on-demand hierarchies may address sensing needs for time-critical ap plications such as autonomous driving.
****

Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference
Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, Long Quan; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2019, pp. 5525-5534
Deep learning has recently demonstrated its excellent performance for multi-view stereo (MVS). However, one major limitation of current learned MVS approaches i s the scalability: the memory-consuming cost volume regularization makes the lea rned MVS hard to be applied to high-resolution scenes. In this paper, we introdu ce a scalable multi-view stereo framework based on the recurrent neural network. Instead of regularizing the entire 3D cost volume in one go, the proposed Recur rent Multi-view Stereo Network (R-MVSNet) sequentially regularizes the 2D cost m aps along the depth direction via the gated recurrent unit (GRU). This reduces d ramatically the memory consumption and makes high-resolution reconstruction feas ible. We first show the state-of-the-art performance achieved by the proposed R- MVSNet on the recent MVS benchmarks. Then, we further demonstrate the scalabilit y of the proposed method on several large-scale scenarios, where previous learne d approaches often fail due to the memory constraint. Code is available at https ://github.com/YoYo000/MVSNet.
****

Synthesizing 3D Shapes From Silhouette Image Collections Using Multi-Projection Generative Adversarial Networks
Xiao Li, Yue Dong, Pieter Peers, Xin Tong; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5535-5544
We present a new weakly supervised learning-based method for generating novel ca tegory-specific 3D shapes from unoccluded image collections. Our method is weakl y supervised and only requires silhouette annotations from unoccluded, category- specific objects. Our method does not require access to the object's 3D shape, m ultiple observations per object from different views, intra-image pixel correspo ndences, or any view annotations. Key to our method is a novel multi-projection generative adversarial network (MP-GAN) that trains a 3D shape generator to be c

onsistent with multiple 2D projections of the 3D shapes, and without direct acce ss to these 3D shapes. This is achieved through multiple discriminators that enc ode the distribution of 2D projections of the 3D shapes seen from a different vi ews. Additionally, to determine the view information for each silhouette image, we also train a view prediction network on visualizations of 3D shapes synthesiz ed by the generator. We iteratively alternate between training the generator and training the view prediction network. We validate our multi-projection GAN on b oth synthetic and real image datasets. Furthermore, we also show that multi-proj ection GANs can aid in learning other high-dimensional distributions from lower dimensional training datasets, such as material-class specific spatially varying reflectance properties from images.
****

The Perfect Match: 3D Point Cloud Matching With Smoothed Densities
Zan Gojcic, Caifa Zhou, Jan D. Wegner, Andreas Wieser; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 55 45-5554
We propose 3DSmoothNet, a full workflow to match 3D point clouds with a siamese deep learning architecture and fully convolutional layers using a voxelized smoo thed density value (SDV) representation. The latter is computed per interest poi nt and aligned to the local reference frame (LRF) to achieve rotation invariance . Our compact, learned, rotation invariant 3D point cloud descriptor achieves 94 .9% average recall on the 3DMatch benchmark data set, outperforming the state-of -the-art by more than 20 percent points with only 32 output dimensions. This ver y low output dimension allows for near realtime correspondence search with 0.1 m s per feature point on a standard PC. Our approach is sensor- and scene-agnostic because of SDV, LRF and learning highly descriptive features with fully convolu tional layers. We show that 3DSmoothNet trained only on RGB-D indoor scenes of b uildings achieves 79.0% average recall on laser scans of outdoor vegetation, mor e than double the performance of our closest, learning-based competitors. Code, data and pre-trained models are available online at https://github.com/zgojcic/3 DSmoothNet.
****

Recurrent Neural Network for (Un-)Supervised Learning of Monocular Video Visual Odometry and Depth
Rui Wang, Stephen M. Pizer, Jan-Michael Frahm; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5555-5564
Deep learning-based, single-view depth estimation methods have recently shown hi ghly promising results. However, such methods ignore one of the most important f eatures for determining depth in the human vision system, which is motion. We pr opose a learning-based, multi-view dense depth map and odometry estimation metho d that uses Recurrent Neural Networks (RNN) and trains utilizing multi-view imag e reprojection and forward-backward flow-consistency losses. Our model can be tr ained in a supervised or even unsupervised mode. It is designed for depth and v isual odometry estimation from video where the input frames are temporally corre lated. However, it also generalizes to single-view depth estimation. Our method produces superior results to the state-of-the-art approaches for single-view an d multi-view learning-based depth estimation on the KITTI driving dataset.
****

PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing
Hengshuang Zhao, Li Jiang, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5565-5 573
This paper presents PointWeb, a new approach to extract contextual features from local neighborhood in a point cloud. Unlike previous work, we densely connect e ach point with every other in a local neighborhood, aiming to specify feature of each point based on the local region characteristics for better representing th e region. A novel module, namely Adaptive Feature Adjustment (AFA) module, is pr esented to find the interaction between points. For each local region, an impact map carrying element-wise impact between point pairs is applied to the feature difference map. Each feature is then pulled or pushed by other features in the s

ame region according to the adaptively learned impact indicators. The adjusted f eatures are well encoded with region information, and thus benefit the point clo ud recognition tasks, such as point cloud segmentation and classification. Exper imental results show that our model outperforms the state-of-the-arts on both se mantic segmentation and shape classification datasets.
****

Scan2Mesh: From Unstructured Range Scans to 3D Meshes
Angela Dai, Matthias Niessner; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2019, pp. 5574-5583
We introduce Scan2Mesh, a novel data-driven generative approach which transforms an unstructured and potentially incomplete range scan into a structured 3D mesh representation. The main contribution of this work is a generative neural netwo rk architecture whose input is a range scan of a 3D object and whose output is a n indexed face set conditioned on the input scan. In order to generate a 3D mesh as a set of vertices and face indices, the generative model builds on a series of proxy losses for vertices, edges, and faces. At each stage, we realize a one-to-one discrete mapping between the predicted and ground truth data points with a combination of convolutional- and graph neural network architectures. This ena bles our algorithm to predict a compact mesh representation similar to those cre ated through manual artist effort using 3D modeling software. Our generated mesh results thus produce sharper, cleaner meshes with a fundamentally different str ucture from those generated through implicit functions, a first step in bridging the gap towards artist-created CAD models.
****

Unsupervised Domain Adaptation for ToF Data Denoising With Adversarial Learning
Gianluca Agresti, Henrik Schaefer, Piergiorgio Sartor, Pietro Zanuttigh; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2019, pp. 5584-5593
Time-of-Flight data is typically affected by a high level of noise and by artifa cts due to Multi-Path Interference (MPI). While various traditional approaches f or ToF data improvement have been proposed, machine learning techniques have sel dom been applied to this task, mostly due to the limited availability of real wo rld training data with depth ground truth. In this paper, we avoid to rely on l abeled real data in the learning framework. A Coarse-Fine CNN, able to exploit m ulti-frequency ToF data for MPI correction, is trained on synthetic data with gr ound truth in a supervised way. In parallel, an adversarial learning strategy, b ased on the Generative Adversarial Networks (GAN) framework, is used to perform an unsupervised pixel-level domain adaptation from synthetic to real world data , exploiting unlabeled real world acquisitions. Experimental results demonstrat e that the proposed approach is able to effectively denoise real world data and to outperform state-of-the-art techniques.
****

Learning Independent Object Motion From Unlabelled Stereoscopic Videos
Zhe Cao, Abhishek Kar, Christian Hane, Jitendra Malik; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 55 94-5603
We present a system for learning motion maps of independently moving objects fro m stereo videos. The only annotations used in our system are 2D object bounding boxes which introduce the notion of objects in our system. Unlike prior learning based approaches which have focused on predicting dense optical flow fields and /or depth maps for images, we propose to predict instance specific 3D scene flow maps and instance masks from which we derive a factored 3D motion map for each object instance. Our network takes the 3D geometry of the problem into account w hich allows it to correlate the input images and distinguish moving objects from static ones. We present experiments evaluating the accuracy of our 3D flow vect ors, as well as depth maps and projected 2D optical flow where our jointly learn ed system outperforms earlier approaches trained for each task independently.
****

Learning Single-Image Depth From Videos Using Quality Assessment Networks
Weifeng Chen, Shengyi Qian, Jia Deng; Proceedings of the IEEE/CVF Conference o

n Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5604-5613
Depth estimation from a single image in the wild remains a challenging problem. One main obstacle is the lack of high-quality training data for images in the wild. In this paper we propose a method to automatically generate such data through Structure-from-Motion (SfM) on Internet videos. The core of this method is a Quality Assessment Network that identifies high-quality reconstructions obtained from SfM. Using this method, we collect single-view depth training data from a large number of YouTube videos and construct a new dataset called YouTube3D. Experiments show that YouTube3D is useful in training depth estimation networks and advances the state of the art of single-view depth estimation in the wild.
****

Learning 3D Human Dynamics From Video
Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5614-5623
From an image of a person in action, we can easily guess the 3D motion of the person in the immediate past and future. This is because we have a mental model of 3D human dynamics that we have acquired from observing visual sequences of humans in motion. We present a framework that can similarly learn a representation of 3D dynamics of humans from video via a simple but effective temporal encoding of image features. At test time, from video, the learned temporal representation give rise to smooth 3D mesh predictions. From a single image, our model can recover the current 3D mesh as well as its 3D past and future motion. Our approach is designed so it can learn from videos with 2D pose annotations in a semi-supervised manner. Though annotated data is always limited, there are millions of videos uploaded daily on the Internet. In this work, we harvest this Internet-scale source of unlabeled data by training our model on unlabeled video with pseudo-ground truth 2D pose obtained from an off-the-shelf 2D pose detector. Our experiments show that adding more videos with pseudo-ground truth 2D pose monotonically improves 3D prediction performance. We evaluate our model on the recent challenging dataset of 3D Poses in the Wild and obtain state-of-the-art performance on the 3D prediction task without any fine-tuning. The project website with video can be found at https://akanazawa.github.io/human_dynamics/.
****

Lending Orientation to Neural Networks for Cross-View Geo-Localization
Liu Liu, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5624-5633
This paper studies image-based geo-localization (IBL) problem using ground-to-aerial cross-view matching. The goal is to predict the spatial location of a ground-level query image by matching it to a large geotagged aerial image database (e.g., satellite imagery). This is a challenging task due to the drastic differences in their viewpoints and visual appearances. Existing deep learning methods for this problem have been focused on maximizing feature similarity between spatially close-by image pairs, while minimizing other images pairs which are far apart. They do so by deep feature embedding based on visual appearance in those ground-and-aerial images. However, in everyday life, humans commonly use orientation information as an important cue for the task of spatial localization. Inspired by this insight, this paper proposes a novel method which endows deep neural networks with the `commonsense' of orientation. Given a ground-level spherical panoramic image as query input (and a large georeferenced satellite image database), we design a Siamese network which explicitly encodes the orientation (i.e., spherical directions) of each pixel of the images. Our method significantly boosts the discriminative power of the learned deep features, leading to a much higher recall and precision outperforming all previous methods. Our network is also more compact using only 1/5th number of parameters than a previously best-performing network. To evaluate the generalization of our method, we also created a large-scale cross-view localization benchmark containing 100K geotagged ground-aerial pairs covering a city. Our codes and datasets are available at https://github.com/Liumouliu/OriCNN.
****

Visual Localization by Learning Objects-Of-Interest Dense Match Regression
Philippe Weinzaepfel, Gabriela Csurka, Yohann Cabon, Martin Humenberger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5634-5643
We introduce a novel CNN-based approach for visual localization from a single RGB image that relies on densely matching a set of Objects-of-Interest (OOIs). In this paper, we focus on planar objects which are highly descriptive in an environment, such as paintings in museums or logos and storefronts in malls or airports. For each OOI, we define a reference image for which 3D world coordinates are available. Given a query image, our CNN model detects the OOIs, segments them and finds a dense set of 2D-2D matches between each detected OOI and its corresponding reference image. Given these 2D-2D matches, together with the 3D world coordinates of each reference image, we obtain a set of 2D-3D matches from which solving a Perspective-n-Point problem gives a pose estimate. We show that 2D-3D matches for reference images, as well as OOI annotations can be obtained for all training images from a single instance annotation per OOI by leveraging Structure-from-Motion reconstruction. We introduce a novel synthetic dataset, VirtualGallery, which targets challenges such as varying lighting conditions and different occlusion levels. Our results show that our method achieves high precision and is robust to these challenges. We also experiment using the Baidu localization dataset captured in a shopping mall. Our approach is the first deep regression-based method to scale to such a larger environment.
****
Bilateral Cyclic Constraint and Adaptive Regularization for Unsupervised Monocular Depth Prediction
Alex Wong, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5644-5653
Supervised learning methods to infer (hypothesize) depth of a scene from a single image require costly per-pixel ground-truth. We follow a geometric approach that exploits abundant stereo imagery to learn a model to hypothesize scene structure without direct supervision. Although we train a network with stereo pairs, we only require a single image at test time to hypothesize disparity or depth. We propose a novel objective function that exploits the bilateral cyclic relationship between the left and right disparities and we introduce an adaptive regularization scheme that allows the network to handle both the co-visible and occluded regions in a stereo pair. This process ultimately produces a model to generate hypotheses for the 3-dimensional structure of the scene as viewed in a single image. When used to generate a single (most probable) estimate of depth, our method outperforms state-of-the-art unsupervised monocular depth prediction methods on the KITTI benchmarks. We show that our method generalizes well by applying our models trained on KITTI to the Make3d dataset.
****
Face Parsing With RoI Tanh-Warping
Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, Lu Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5654-5663
Face parsing computes pixel-wise label maps for different semantic components (e.g., hair, mouth, eyes) from face images. Existing face parsing literature have illustrated significant advantages by focusing on individual regions of interest (RoIs) for faces and facial components. However,the traditional crop-and-resize focusing mechanism ignores all contextual area outside the RoIs, and thus is not suitable when the component area is unpredictable, e.g. hair. Inspired by the physiological vision system of human, we propose a novel RoI Tanh-warping operator that combines the central vision and the peripheral vision together. It addresses the dilemma between a limited sized RoI for focusing and an unpredictable area of surrounding context for peripheral information. To this end, we propose a novel hybrid convolutional neural network for face parsing. It uses hierarchical local based method for inner facial components and global methods for outer facial components. The whole framework is simple and principled, and can be trained end-to-end. To facilitate future research of face parsing, we also manually re

label the training data of the HELEN dataset and will make it public. Experiments on both HELEN and LFW-PL benchmarks demonstrate that our method surpasses state-of-the-art methods.
****

Multi-Person Articulated Tracking With Spatial and Temporal Embeddings
Sheng Jin, Wentao Liu, Wanli Ouyang, Chen Qian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5664-5673
We propose a unified framework for multi-person pose estimation and tracking. Our framework consists of two main components, i.e. SpatialNet and TemporalNet. The SpatialNet accomplishes body part detection and part-level data association in a single frame, while the TemporalNet groups human instances in consecutive frames into trajectories. Specifically, besides body part detection heatmaps, SpatialNet also predicts the Keypoint Embedding (KE) and Spatial Instance Embedding (SIE) for body part association. We model the grouping procedure into a differentiable Pose-Guided Grouping (PGG) module to make the whole part detection and grouping pipeline fully end-to-end trainable. TemporalNet extends the spatial grouping of keypoints to temporal grouping of human instances. Given human proposals from two consecutive frames, TemporalNet exploits both appearance features encoded in Human Embedding (HE) and temporally consistent geometric features embodied in Temporal Instance Embedding (TIE) for robust tracking. Extensive experiments demonstrate the effectiveness of our proposed model. Remarkably, we demonstrate substantial improvements over the state-of-the-art pose tracking method from 65.4% to 71.8% Multi-Object Tracking Accuracy (MOTA) on the ICCV'17 PoseTrack Dataset.
****

Multi-Person Pose Estimation With Enhanced Channel-Wise and Spatial Information
Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, Changhu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5674-5682
Multi-person pose estimation is an important but challenging problem in computer vision. Although current approaches have achieved significant progress by fusing the multi-scale feature maps, they pay little attention to enhancing the channel-wise and spatial information of the feature maps. In this paper, we propose two novel modules to perform the enhancement of the information for the multi-person pose estimation. First, a Channel Shuffle Module (CSM) is proposed to adopt the channel shuffle operation on the feature maps with different levels, promoting cross-channel information communication among the pyramid feature maps. Second, a Spatial, Channel-wise Attention Residual Bottleneck (SCARB) is designed to boost the original residual unit with attention mechanism, adaptively highlighting the information of the feature maps both in the spatial and channel-wise context. The effectiveness of our proposed modules is evaluated on the COCO keypoint benchmark, and experimental results show that our approach achieves the state-of-the-art results.
****

A Compact Embedding for Facial Expression Similarity
Raviteja Vemulapalli, Aseem Agarwala; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5683-5692
Most of the existing work on automatic facial expression analysis focuses on discrete emotion recognition, or facial action unit detection. However, facial expressions do not always fall neatly into pre-defined semantic categories. Also, the similarity between expressions measured in the action unit space need not correspond to how humans perceive expression similarity. Different from previous work, our goal is to describe facial expressions in a continuous fashion using a compact embedding space that mimics human visual preferences. To achieve this goal, we collect a large-scale faces-in-the-wild dataset with human annotations in the form: Expressions A and B are visually more similar when compared to expression C, and use this dataset to train a neural network that produces a compact (16-dimensional) expression embedding. We experimentally demonstrate that the learned embedding can be successfully used for various applications such as expression retrieval, photo album summarization, and emotion recognition. We also show th

at the embedding learned using the proposed dataset performs better than several other embeddings learned using existing emotion or action unit datasets.
****

Deep High-Resolution Representation Learning for Human Pose Estimation
Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5693-5703
In this paper, we are interested in the human pose estimation problem with a focus on learning reliable high-resolution representations. Most existing methods recover high-resolution representations from low-resolution representations produced by a high-to-low resolution network. Instead, our proposed network maintains high-resolution representations through the whole process. We start from a high-resolution subnetwork as the first stage, gradually add high-to-low resolution subnetworks one by one to form more stages, and connect the mutli-resolution subnetworks in parallel. We conduct repeated multi-scale fusions such that each of the high-to-low resolution representations receives information from other parallel representations over and over, leading to rich high-resolution representations. As a result, the predicted keypoint heatmap is potentially more accurate and spatially more precise. We empirically demonstrate the effectiveness of our network through the superior pose estimation results over two benchmark datasets: the COCO keypoint detection dataset and the MPII Human Pose dataset. In addition, we show the superiority of our network in pose tracking on the PoseTrack dataset. The code and models have been publicly available at https://github.com/leoxiaobin/deep-high-resolution-net.pytorch.
****

Feature Transfer Learning for Face Recognition With Under-Represented Data
Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5704-5713
Despite the large volume of face recognition datasets, there is a significant portion of subjects, of which the samples are insufficient and thus under-represented. Ignoring such significant portion results in insufficient training data. Training with under-represented data leads to biased classifiers in conventionally-trained deep networks. In this paper, we propose a center-based feature transfer framework to augment the feature space of under-represented subjects from the regular subjects that have sufficiently diverse samples. A Gaussian prior of the variance is assumed across all subjects and the variance from regular ones are transferred to the under-represented ones. This encourages the under-represented distribution to be closer to the regular distribution. Further, an alternating training regimen is proposed to simultaneously achieve less biased classifiers and a more discriminative feature representation. We conduct ablative study to mimic the under-represented datasets by varying the portion of under-represented classes on the MS-Celeb-1M dataset. Advantageous results on LFW, IJB-A and MS-Celeb-1M demonstrate the effectiveness of our feature transfer and training strategy, compared to both general baselines and state-of-the-art methods. Moreover, our feature transfer successfully presents smooth visual interpolation, which conducts disentanglement to preserve identity of a class while augmenting its feature space with non-identity variations such as pose and lighting.
****

Unsupervised 3D Pose Estimation With Geometric Self-Supervision
Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, James M. Rehg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5714-5724
We present an unsupervised learning approach to re- cover 3D human pose from 2D skeletal joints extracted from a single image. Our method does not require any multi- view image data, 3D skeletons, correspondences between 2D-3D points, or use previously learned 3D priors during training. A lifting network accepts 2D landmarks as inputs and generates a corresponding 3D skeleton estimate. Dur- ing training, the recovered 3D skeleton is reprojected on random camera viewpoints to generate new 'synthetic' 2D poses. By lifting the synthetic 2D poses back to 3D and re-projecting them in the original camera view, we can de- fine self-consist

ency loss both in 3D and in 2D. The training can thus be self supervised by expl oiting the geometric self- consistency of the lift-reproject-lift process. We sh ow that self-consistency alone is not sufficient to generate realistic skeletons , however adding a 2D pose discriminator enables the lifter to output valid 3D p oses. Additionally, to learn from 2D poses 'in the wild', we train an unsupervis ed 2D domain adapter network to allow for an expansion of 2D data. This improves results and demonstrates the useful- ness of 2D pose data for unsupervised 3D l ifting. Results on Human3.6M dataset for 3D human pose estimation demon- strate that our approach improves upon the previous un- supervised methods by 30% and o utperforms many weakly supervised approaches that explicitly use 3D data.
****

Peeking Into the Future: Predicting Future Person Activities and Locations in Vi deos

Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G. Hauptmann, Li Fei- Fei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 5725-5734

Deciphering human behaviors to predict their future paths/trajectories and what they would do from videos is important in many applications. Motivated by this i dea, this paper studies predicting a pedestrian's future path jointly with futur e activities. We propose an end-to-end, multi-task learning system utilizing ric h visual features about human behavioral information and interaction with their surroundings. To facilitate the training, the network is learned with an auxilia ry task of predicting future location in which the activity will happen. Experim ental results demonstrate our state-of-the-art performance over two public bench marks on future trajectory prediction. Moreover, our method is able to produce m eaningful future activity prediction in addition to the path. The result provide s the first empirical evidence that joint modeling of paths and activities benef its future path prediction.
****

Re-Identification With Consistent Attentive Siamese Networks

Meng Zheng, Srikrishna Karanam, Ziyan Wu, Richard J. Radke; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p p. 5735-5744

We propose a new deep architecture for person re-identification (re-id). While r e-id has seen much recent progress, spatial localization and view-invariant repr esentation learning for robust cross-view matching remain key, unsolved problems . We address these questions by means of a new attention-driven Siamese learning architecture, called the Consistent Attentive Siamese Network. Our key innovati ons compared to existing, competing methods include (a) a flexible framework des ign that produces attention with only identity labels as supervision, (b) explic it mechanisms to enforce attention consistency among images of the same person, and (c) a new Siamese framework that integrates attention and attention consiste ncy, producing principled supervisory signals as well as the first mechanism tha t can explain the reasoning behind the Siamese framework's predictions. We condu ct extensive evaluations on the CUHK03-NP, DukeMTMC-ReID, and Market-1501 datase ts and report competitive performance.
****

On the Continuity of Rotation Representations in Neural Networks

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, Hao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp . 5745-5753

In neural networks, it is often desirable to work with various representations o f the same space. For example, 3D rotations can be represented with quaternions or Euler angles. In this paper, we advance a definition of a continuous represe ntation, which can be helpful for training deep neural networks. We relate this to topological concepts such as homeomorphism and embedding. We then investigate what are continuous and discontinuous representations for 2D, 3D, and n-dimensi onal rotations. We demonstrate that for 3D rotations, all representations are di scontinuous in the real Euclidean spaces of four or fewer dimensions. Thus, wide ly used representations such as quaternions and Euler angles are discontinuous a

nd difficult for neural networks to learn. We show that the 3D rotations have co ntinuous representations in 5D and 6D, which are more suitable for learning. We also present continuous representations for the general case of the n-dimensiona l rotation group SO(n). While our main focus is on rotations, we also show that our constructions apply to other groups such as the orthogonal group and similar ity transforms. We finally present empirical results, which show that our contin uous rotation representations outperform discontinuous ones for several practica l problems in graphics and vision, including a simple autoencoder sanity test, a rotation estimator for 3D point clouds, and an inverse kinematics solver for 3D human poses.
****

Iterative Residual Refinement for Joint Optical Flow and Occlusion Estimation
Junhwa Hur, Stefan Roth; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2019, pp. 5754-5763
Deep learning approaches to optical flow estimation have seen rapid progress ove r the recent years. One common trait of many networks is that they refine an ini tial flow estimate either through multiple stages or across the levels of a coar se-to-fine representation. While leading to more accurate results, the downside of this is an increased number of parameters. Taking inspiration from both class ical energy minimization approaches as well as residual networks, we propose an iterative residual refinement (IRR) scheme based on weight sharing that can be c ombined with several backbone networks. It reduces the number of parameters, imp roves the accuracy, or even achieves both. Moreover, we show that integrating oc clusion prediction and bi-directional flow estimation into our IRR scheme can fu rther boost the accuracy. Our full network achieves state-of-the-art results for both optical flow and occlusion estimation across several standard datasets.
****

Inverse Discriminative Networks for Handwritten Signature Verification
Ping Wei, Huan Li, Ping Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5764-5772
Handwritten signature verification is an important technique for many financial, commercial, and forensic applications. In this paper, we propose an inverse dis criminative network (IDN) for writer-independent handwritten signature verificat ion, which aims to determine whether a test signature is genuine or forged compa red to the reference signature. The IDN model contains four weight-shared neural network streams, of which two receiving the original signature images are the d iscriminative streams and the other two addressing the gray-inverted images form the inverse streams. Multiple paths of attention modules connect the discrimina tive streams and the inverse streams to propagate messages. With the inverse str eams and the multi-path attention modules, the IDN model intensifies the effecti ve information of signature verification. Since there was no proper Chinese sign ature dataset in the community, we collected a large-scale Chinese signature dat aset with approximately 29,000 images of 749 individuals' signatures. We test ou r method on the Chinese signature dataset and other three signature datasets of different languages: CEDAR, BHSig-B, and BHSig-H. Experiments prove the strength and potential of our method.
****

Led3D: A Lightweight and Efficient Deep Approach to Recognizing Low-Quality 3D F aces
Guodong Mu, Di Huang, Guosheng Hu, Jia Sun, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp . 5773-5782
Due to the intrinsic invariance to pose and illumination changes, 3D Face Recogn ition (FR) has a promising potential in the real world. 3D FR using high-quality faces, which are of high resolutions and with smooth surfaces, have been widely studied. However, research on that with low-quality input is limited, although it involves more applications. In this paper, we focus on 3D FR using low-qualit y data, targeting an efficient and accurate deep learning solution. To achieve t his, we work on two aspects: (1) designing a lightweight yet powerful CNN; (2) g enerating finer and bigger training data. For (1), we propose a Multi-Scale Feat

ure Fusion (MSFF) module and a Spatial Attention Vectorization (SAV) module to b
uild a compact and discriminative CNN. For (2), we propose a data processing sys
tem including point-cloud recovery, surface refinement, and data augmentation (w
ith newly proposed shape jittering and shape scaling). We conduct extensive expe
riments on Lock3DFace and achieve state-of-the-art results, outperforming many h
eavy CNNs such as VGG-16 and ResNet-34. In addition, our model can operate at a
very high speed (136 fps) on Jetson TX2, and the promising accuracy and efficien
cy reached show its great applicability on edge/mobile devices.
****

ROI Pooled Correlation Filters for Visual Tracking
Yuxuan Sun, Chong Sun, Dong Wang, You He, Huchuan Lu; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 57
83-5791
The ROI (region-of-interest) based pooling method performs pooling operations on
the cropped ROI regions for various samples and has shown great success in the
object detection methods. It compresses the model size while preserving the loca
lization accuracy, thus it is useful in the visual tracking field. Though being
effective, the ROI-based pooling operation is not yet considered in the correlat
ion filter formula. In this paper, we propose a novel ROI pooled correlation fil
ter (RPCF) algorithm for robust visual tracking. Through mathematical derivation
s, we show that the ROI-based pooling can be equivalently achieved by enforcing
additional constraints on the learned filter weights, which makes the ROI-based
pooling feasible on the virtual circular samples. Besides, we develop an efficie
nt joint training formula for the proposed correlation filter algorithm, and der
ive the Fourier solvers for efficient model training. Finally, we evaluate our R
PCF tracker on OTB-2013, OTB-2015 and VOT-2017 benchmark datasets. Experimental
results show that our tracker performs favourably against other state-of-the-art
trackers.
****

Deep Video Inpainting
Dahun Kim, Sanghyun Woo, Joon-Young Lee, In So Kweon; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 579
2-5801
Video inpainting aims to fill spatio-temporal holes with plausible content in a
video. Despite tremendous progress of deep neural networks for image inpainting,
it is challenging to extend these methods to the video domain due to the additi
onal time dimension. In this work, we propose a novel deep network architecture
for fast video inpainting. Built upon an image-based encoder-decoder model, our
framework is designed to collect and refine information from neighbor frames and
synthesize still-unknown regions. At the same time, the output is enforced to b
e temporally consistent by a recurrent feedback and a temporal memory module. Co
mpared with the state-of-the-art image inpainting algorithm, our method produces
videos that are much more semantically correct and temporally smooth. In contra
st to the prior video completion method which relies on time-consuming optimizat
ion, our method runs in near real-time while generating competitive video result
s. Finally, we applied our framework to video retargeting task, and obtain visua
lly pleasing results.
****

DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthes
is
Minfeng Zhu, Pingbo Pan, Wei Chen, Yi Yang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5802-5810
In this paper, we focus on generating realistic images from text descriptions. C
urrent methods first generate an initial image with rough shape and color, and t
hen refine the initial image to a high-resolution one. Most existing text-to-ima
ge synthesis methods have two main problems. (1) These methods depend heavily on
the quality of the initial images. If the initial image is not well initialized
, the following processes can hardly refine the image to a satisfactory quality.
(2) Each word contributes a different level of importance when depicting differ
ent image contents, however, unchanged text representation is used in existing i

mage refinement processes. In this paper, we propose the Dynamic Memory Generati ve Adversarial Network (DM-GAN) to generate high-quality images. The proposed me thod introduces a dynamic memory module to refine fuzzy image contents, when the initial images are not well generated. A memory writing gate is designed to sel ect the important text information based on the initial image content, which ena bles our method to accurately generate images from the text description. We also utilize a response gate to adaptively fuse the information read from the memori es and the image features. We evaluate the DM-GAN model on the Caltech-UCSD Bird s 200 dataset and the Microsoft Common Objects in Context dataset. Experimental results demonstrate that our DM-GAN model performs favorably against the state-o f-the-art approaches.

****

Non-Adversarial Image Synthesis With Generative Latent Nearest Neighbors

Yedid Hoshen, Ke Li, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5811-5819

Unconditional image generation has recently been dominated by generative adversa rial networks (GANs). GAN methods train a generator which regresses images from random noise vectors, as well as a discriminator that attempts to differentiate between the generated images and a training set of real images. GANs have shown amazing results at generating realistic looking images. Despite their success, G ANs suffer from critical drawbacks including: unstable training and mode-droppin g. The weaknesses in GANs have motivated research into alternatives including: v ariational auto-encoders (VAEs), latent embedding learning methods (e.g. GLO) an d nearest-neighbor based implicit maximum likelihood estimation (IMLE). Unfortun ately at the moment, GANs still significantly outperform the alternative methods for image generation. In this work, we present a novel method - Generative Late nt Nearest Neighbors (GLANN) - for training generative models without adversaria l training. GLANN combines the strengths of IMLE and GLO in a way that overcomes the main drawbacks of each method. Consequently, GLANN generates images that ar e far better than GLO and IMLE. Our method does not suffer from mode collapse wh ich plagues GAN training and is much more stable. Qualitative results show that GLANN outperforms a baseline consisting of 800 GANs and VAEs on commonly used da tasets. Our models are also shown to be effective for training truly non-adversa rial unsupervised image translation.

****

Mixture Density Generative Adversarial Networks

Hamid Eghbal-zadeh, Werner Zellinger, Gerhard Widmer; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5820 -5829

Generative Adversarial Networks have a surprising ability to generate sharp and realistic images, but they are known to suffer from the so-called mode collapse problem. In this paper, we propose a new GAN variant called Mixture Density GAN that overcomes this problem by encouraging the Discriminator to form clusters in its embedding space, which in turn leads the Generator to exploit these and dis cover different modes in the data. This is achieved by positioning Gaussian dens ity functions in the corners of a simplex, using the resulting Gaussian mixture as a likelihood function over discriminator embeddings, and formulating an objec tive function for GAN training that is based on these likelihoods. We show how f ormation of these clusters changes the probability landscape of the discriminato r and improves the mode discovery of the GAN. We also show that the optimum of o ur training objective is attained if and only if the generated and the real dist ribution match exactly. We support our theoretical results with empirical evalua tions on three mode discovery benchmark datasets (Stacked-MNIST, Ring of Gaussia ns and Grid of Gaussians), and four image datasets (CIFAR-10, CelebA, MNIST, and Fashion-MNIST). Furthermore, we demonstrate (1) the ability to avoid mode colla pse and discover all the modes and (2) superior quality of the generated images (as measured by the Frechet Inception Distance (FID)), achieving the lowest FID compared to all baselines.

****

SketchGAN: Joint Sketch Completion and Recognition With Generative Adversarial N

etwork
Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, Hongan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2019, pp. 5830-5839
Hand-drawn sketch recognition is a fundamental problem in computer vision, widel y used in sketch-based image and video retrieval, editing, and reorganization. P revious methods often assume that a complete sketch is used as input; however, h and-drawn sketches in common application scenarios are often incomplete, which m akes sketch recognition a challenging problem. In this paper, we propose SketchG AN, a new generative adversarial network (GAN) based approach that jointly compl etes and recognizes a sketch, boosting the performance of both tasks. Specifical ly, we use a cascade Encode-Decoder network to complete the input sketch in an i terative manner, and employ an auxiliary sketch recognition task to recognize th e completed sketch. Experiments on the Sketchy database benchmark demonstrate th at our joint learning approach achieves competitive sketch completion and recogn ition performance compared with the state-of-the-art methods. Further experiment s using several sketch-based applications also validate the performance of our m ethod.
****

Foreground-Aware Image Inpainting
Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2019, pp. 5840-5848
Existing image inpainting methods typically fill holes by borrowing information from surrounding pixels. They often produce unsatisfactory results when the hole s overlap with or touch foreground objects due to lack of information about the actual extent of foreground and background regions within the holes. These scena rios, however, are very important in practice, especially for applications such as distracting object removal. To address the problem, we propose a foreground-a ware image inpainting system that explicitly disentangles structure inference an d content completion. Specifically, our model learns to predict the foreground c ontour first, and then inpaints the missing region using the predicted contour a s guidance. We show that by such disentanglement, the contour completion model p redicts reasonable contours of objects, and further substantially improves the p erformance of image inpainting. Experiments show that our method significantly o utperforms existing methods and achieves superior inpainting results on challeng ing cases with complex compositions.
****

Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-To-Imag e Translation
Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 201 9, pp. 5849-5859
The applicability of computer vision to real paintings and artworks has been rar ely investigated, even though a vast heritage would greatly benefit from techniq ues which can understand and process data from the artistic domain. This is part ially due to the small amount of annotated artistic data, which is not even comp arable to that of natural images captured by cameras. In this paper, we propose a semantic-aware architecture which can translate artworks to photo-realistic vi sualizations, thus reducing the gap between visual features of artistic and real istic data. Our architecture can generate natural images by retrieving and learn ing details from real photos through a similarity matching strategy which lever ages a weakly-supervised semantic understanding of the scene. Experimental resul ts show that the proposed technique leads to increased realism and to a reductio n in domain shift, which improves the performance of pre-trained architectures f or classification, detection, and segmentation. Code is publicly available at: h ttps://github.com/aimagelab/art2real.
****

Structure-Preserving Stereoscopic View Synthesis With Multi-Scale Adversarial Co rrelation Matching

Yu Zhang, Dongqing Zou, Jimmy S. Ren, Zhe Jiang, Xiaohao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5860-5869

This paper addresses stereoscopic view synthesis from a single image. Various recent works solve this task by reorganizing pixels from the input view to reconstruct the target one in a stereo setup. However, purely depending on such photometric-based reconstruction process, the network may produce structurally inconsistent results. Regarding this issue, this work proposes Multi-Scale Adversarial Correlation Matching (MS-ACM), a novel learning framework for structure-aware view synthesis. The proposed framework does not assume any costly supervision signal of scene structures such as depth. Instead, it models structures as self-correlation coefficients extracted from multi-scale feature maps in transformed spaces. In training, the feature space attempts to push the correlation distances between the synthesized and target images far apart, thus amplifying inconsistent structures. At the same time, the view synthesis network minimizes such correlation distances by fixing mistakes it makes. With such adversarial training, structural errors of different scales and levels are iteratively discovered and reduced, preserving both global layouts and fine-grained details. Extensive experiments on the KITTI benchmark show that MS-ACM improves both visual quality and the metrics over existing methods when plugged into recent view synthesis architectures.

****

DynTypo: Example-Based Dynamic Text Effects Transfer

Yifang Men, Zhouhui Lian, Yingmin Tang, Jianguo Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5870-5879

In this paper, we present a novel approach for dynamic text effects transfer by using example-based texture synthesis. In contrast to previous works that require an input video of the target to provide motion guidance, we aim to animate a still image of the target text by transferring the desired dynamic effects from an observed exemplar. Due to the simplicity of target guidance and complexity of realistic effects, it is prone to producing temporal artifacts such as flickers and pulsations. To address the problem, our core idea is to find a common Nearest-neighbor Field (NNF) that would optimize the textural coherence across all key frames simultaneously. With the static NNF for video sequences, we implicitly transfer motion properties from source to target. We also introduce a guided NNF search by employing the distance-based weight map and Simulated Annealing (SA) for deep direction-guided propagation to allow intense dynamic effects to be completely transferred with no semantic guidance provided. Experimental results demonstrate the effectiveness and superiority of our method in dynamic text effects transfer through extensive comparisons with state-of-the-art algorithms. We also show the potentiality of our method via multiple experiments for various application domains.

****

Arbitrary Style Transfer With Style-Attentional Networks

Dae Young Park, Kwang Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5880-5888

Arbitrary style transfer aims to synthesize a content image with the style of an image to create a third image that has never been seen before. Recent arbitrary style transfer algorithms find it challenging to balance the content structure and the style patterns. Moreover, simultaneously maintaining the global and local style patterns is difficult due to the patch-based mechanism. In this paper, we introduce a novel style-attentional network (SANet) that efficiently and flexibly integrates the local style patterns according to the semantic spatial distribution of the content image. A new identity loss function and multi-level feature embeddings enable our SANet and decoder to preserve the content structure as much as possible while enriching the style patterns. Experimental results demonstrate that our algorithm synthesizes stylized images in real-time that are higher in quality than those produced by the state-of-the-art algorithms.

****

Typography With Decor: Intelligent Text Style Transfer
Wenjing Wang, Jiaying Liu, Shuai Yang, Zongming Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5889-5897
Text effects transfer can dramatically make the text visually pleasing. In this paper, we present a novel framework to stylize the text with exquisite decor, which is ignored by the previous text stylization methods. Decorative elements pose a challenge to spontaneously handle basal text effects and decor, which are two different styles. To address this issue, our key idea is to learn to separate, transfer and recombine the decors and the basal text effect. A novel text effect transfer network is proposed to infer the styled version of the target text. The stylized text is finally embellished with decor where the placement of the decor is carefully determined by a novel structure-aware strategy. Furthermore, we propose a domain adaptation strategy for decor detection and a one-shot training strategy for text effects transfer, which greatly enhance the robustness of our network to new styles. We base our experiments on our collected topography dataset including 59,000 professionally styled text and demonstrate the superiority of our method over other state-of-the-art style transfer methods.
****
RL-GAN-Net: A Reinforcement Learning Agent Controlled GAN Network for Real-Time Point Cloud Shape Completion
Muhammad Sarmad, Hyunjoo Jenny Lee, Young Min Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5898-5907
We present RL-GAN-Net, where a reinforcement learning (RL) agent provides fast and robust control of a generative adversarial network (GAN). Our framework is applied to point cloud shape completion that converts noisy, partial point cloud data into a high-fidelity completed shape by controlling the GAN. While a GAN is unstable and hard to train, we circumvent the problem by (1) training the GAN on the latent space representation whose dimension is reduced compared to the raw point cloud input and (2) using an RL agent to find the correct input to the GAN to generate the latent space representation of the shape that best fits the current input of incomplete point cloud. The suggested pipeline robustly completes point cloud with large missing regions. To the best of our knowledge, this is the first attempt to train an RL agent to control the GAN, which effectively learns the highly nonlinear mapping from the input noise of the GAN to the latent space of point cloud. The RL agent replaces the need for complex optimization and consequently makes our technique real time. Additionally, we demonstrate that our pipelines can be used to enhance the classification accuracy of point cloud with missing data.
****
Photo Wake-Up: 3D Character Animation From a Single Photo
Chung-Yi Weng, Brian Curless, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5908-5917
We present a method and application for animating a human subject from a single photo. E.g., the character can walk out, run, sit, or jump in 3D. The key contributions of this paper are: 1) an application of viewing and animating humans in single photos in 3D, 2) a novel 2D warping method to deform a posable template body model to fit the person's complex silhouette to create an animatable mesh, and 3) a method for handling partial self occlusions. We compare to state-of-the-art related methods and evaluate results with human studies. Further, we present an interactive interface that allows re-posing the person in 3D, and an augmented reality setup where the animated 3D person can emerge from the photo into the real world. We demonstrate the method on photos, posters, and art. The project page is at https://grail.cs.washington.edu/projects/wakeup/.
****
DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality
Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, Paul Debevec; Proceedings of the IEEE/CVF Conference on Computer Vis

ion and Pattern Recognition (CVPR), 2019, pp. 5918-5928
We present a learning-based method to infer plausible high dynamic range (HDR), omnidirectional illumination given an unconstrained, low dynamic range (LDR) image from a mobile phone camera with a limited field of view (FOV). For training data, we collect videos of various reflective spheres placed within the camera's FOV, leaving most of the background unoccluded, leveraging that materials with diverse reflectance functions reveal different lighting cues in a single exposure. We train a deep neural network to regress from the LDR background image to HDR lighting by matching the LDR ground truth sphere images to those rendered with the predicted illumination using image-based relighting, which is differentiable. Our inference runs at interactive frame rates on a mobile device, enabling realistic rendering of virtual objects into real scenes for mobile mixed reality. Training on automatically exposed and white-balanced videos, we improve the realism of rendered objects compared to the state-of-the art methods for both indoor and outdoor scenes.
****

Iterative Residual CNNs for Burst Photography Applications
Filippos Kokkinos, Stamatis Lefkimmiatis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5929-5938
Modern inexpensive imaging sensors suffer from inherent hardware constraints which often result in captured images of poor quality. Among the most common ways to deal with such limitations is to rely on burst photography, which nowadays acts as the backbone of all modern smartphone imaging applications. In this work, we focus on the fact that every frame of a burst sequence can be accurately described by a forward (physical) model. This, in turn, allows us to restore a single image of higher quality from a sequence of low-quality images as the solution of an optimization problem. Inspired by an extension of the gradient descent method that can handle non-smooth functions, namely the proximal gradient descent, and modern deep learning techniques, we propose a convolutional iterative network with a transparent architecture. Our network uses a burst of low-quality image frames and is able to produce an output of higher image quality recovering fine details which are not distinguishable in any of the original burst frames. We focus both on the burst photography pipeline as a whole, i.e., burst demosaicking and denoising, as well as on the traditional Gaussian denoising task. The developed method demonstrates consistent state-of-the art performance across the two tasks and as opposed to other recent deep learning approaches does not have any inherent restrictions either to the number of frames or their ordering.
****

Learning Implicit Fields for Generative Shape Modeling
Zhiqin Chen, Hao Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5939-5948
We advocate the use of implicit fields for learning generative models of shapes and introduce an implicit field decoder, called IM-NET, for shape generation, aimed at improving the visual quality of the generated shapes. An implicit field assigns a value to each point in 3D space, so that a shape can be extracted as an iso-surface. IM-NET is trained to perform this assignment by means of a binary classifier. Specifically, it takes a point coordinate, along with a feature vector encoding a shape, and outputs a value which indicates whether the point is outside the shape or not. By replacing conventional decoders by our implicit decoder for representation learning (via IM-AE) and shape generation (via IM-GAN), we demonstrate superior results for tasks such as generative shape modeling, interpolation, and single-view 3D reconstruction, particularly in terms of visual quality. Code and supplementary material are available at https://github.com/czq142857/implicit-decoder.
****

Reliable and Efficient Image Cropping: A Grid Anchor Based Approach
Hui Zeng, Lida Li, Zisheng Cao, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5949-5957
Image cropping aims to improve the composition as well as aesthetic quality of an image by removing extraneous content from it. Existing image cropping database

s provide only one or several human-annotated bounding boxes as the groundtruth, which cannot reflect the non-uniqueness and flexibility of image cropping in practice. The employed evaluation metrics such as intersection-over-union cannot reliably reflect the real performance of cropping models, either. This work revisits the problem of image cropping, and presents a grid anchor based formulation by considering the special properties and requirements (e.g., local redundancy, content preservation, aspect ratio) of image cropping. Our formulation reduces the searching space of candidate crops from millions to less than one hundred. Consequently, a grid anchor based cropping benchmark is constructed, where all crops of each image are annotated and more reliable evaluation metrics are defined. We also design an effective and lightweight network module, which simultaneously considers the region of interest and region of discard for more accurate image cropping. Our model can stably output visually pleasing crops for images of different scenes and run at a speed of 125 FPS.

****

Patch-Based Progressive 3D Point Set Upsampling

Wang Yifan, Shihao Wu, Hui Huang, Daniel Cohen-Or, Olga Sorkine-Hornung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5958-5967

We present a detail-driven deep neural network for point set upsampling. A high-resolution point set is essential for point-based rendering and surface reconstruction. Inspired by the recent success of neural image super-resolution techniques, we progressively train a cascade of patch-based upsampling networks on different levels of detail end-to-end. We propose a series of architectural design contributions that lead to a substantial performance boost. The effect of each technical contribution is demonstrated in an ablation study. Qualitative and quantitative experiments show that our method significantly outperforms the state-of-the-art learning-based and optimazation-based approaches, both in terms of handling low-resolution inputs and revealing high-fidelity details.

****

An Iterative and Cooperative Top-Down and Bottom-Up Inference Network for Salient Object Detection

Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5968-5977

This paper presents a salient object detection method that integrates both top-down and bottom-up saliency inference in an iterative and cooperative manner. The top-down process is used for coarse-to-fine saliency estimation, where high-level saliency is gradually integrated with finer lower-layer features to obtain a fine-grained result. The bottom-up process infers the high-level, but rough saliency through gradually using upper-layer, semantically-richer features. These two processes are alternatively performed, where the bottom-up process uses the fine-grained saliency obtained from the top-down process to yield enhanced high-level saliency estimate, and the top-down process, in turn, is further benefited from the improved high-level information. The network layers in the bottom-up/top-down processes are equipped with recurrent mechanisms for layer-wise, step-by-step optimization. Thus, saliency information is effectively encouraged to flow in a bottom-up, top-down and intra-layer manner. We show that most other saliency models based on fully convolutional networks (FCNs) are essentially variants of our model. Extensive experiments on several famous benchmarks clearly demonstrate the superior performance, good generalization, and powerful learning ability of our proposed saliency inference framework.

****

Deep Stacked Hierarchical Multi-Patch Network for Image Deblurring

Hongguang Zhang, Yuchao Dai, Hongdong Li, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5978-5986

Despite deep end-to-end learning methods have shown their superiority in removing non-uniform motion blur, there still exist major challenges with the current multi-scale and scale-recurrent models: 1) Deconvolution/upsampling operations i

n the coarse-to-fine scheme result in expensive runtime; 2) Simply increasing th e model depth with finer-scale levels cannot improve the quality of deblurring. To tackle the above problems, we present a deep hierarchical multi-patch networ k inspired by Spatial Pyramid Matching to deal with blurry images via a fine-to -coarse hierarchical representation. To deal with the performance saturation w.r .t. depth, we propose a stacked version of our multi-patch model. Our proposed b asic multi-patch model achieves the state-of-the-art performance on the GoPro da taset while enjoying a 40xfaster runtime compared to current multi-scale methods . With 30ms to process an image at 1280x720 resolution, it is the first real-tim e deep motion deblurring model for 720p images at 30fps. For stacked networks, s ignificant improvements (over 1.2dB) are achieved on the GoPro dataset by increa sing the network depth. Moreover, by varying the depth of the stacked model, one can adapt the performance and runtime of the same network for different applica tion scenarios.

****

Turn a Silicon Camera Into an InGaAs Camera

Feifan Lv, Yinqiang Zheng, Bohan Zhang, Feng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5987-599 5

Short-wave infrared (SWIR) imaging has a wide range of applications for both ind ustry and civilian. However, the InGaAs sensors commonly used for SWIR imaging s uffer from a variety of drawbacks, including high price, low resolution, unstabl e quality, and so on. In this paper, we propose a novel solution for SWIR imagin g using a common Silicon sensor, which has cheaper price, higher resolution and better technical maturity compared with the specialized InGaAs sensor. Our key i dea is to approximate the response of the InGaAs sensor by exploiting the largel y ignored sensitivity of a Silicon sensor, weak as it is, in the SWIR range. To this end, we build a multi-channel optical system to collect a new SWIR dataset and present a physically meaningful three-stage image processing algorithm on th e basis of CNN. Both qualitative and quantitative experiments show promising exp erimental results, which demonstrate the effectiveness of the proposed method.

****

Low-Rank Tensor Completion With a New Tensor Nuclear Norm Induced by Invertible Linear Transforms

Canyi Lu, Xi Peng, Yunchao Wei; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2019, pp. 5996-6004

This work studies the low-rank tensor completion problem, which aims to exactly recover a low-rank tensor from partially observed entries. Our model is inspired by the recently proposed tensor-tensor product (t-product) based on any inverti ble linear transforms. When the linear transforms satisfy certain conditions, we deduce the new tensor tubal rank, tensor spectral norm, and tensor nuclear norm . Equipped with the tensor nuclear norm, we then solve the tensor completion pro blem by solving a convex program and provide the theoretical bound for the exact recovery under certain tensor incoherence conditions. The achieved sampling com plexity is order-wise optimal. Our model and result greatly extend existing resu lts in the low-rank matrix and tensor completion. Numerical experiments verify o ur results and the application on image recovery demonstrates the superiority of our method.

****

Joint Representative Selection and Feature Learning: A Semi-Supervised Approach

Suchen Wang, Jingjing Meng, Junsong Yuan, Yap-Peng Tan; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6 005-6013

In this paper, we propose a semi-supervised approach for representative selectio n, which finds a small set of representatives that can well summarize a large da ta collection. Given labeled source data and big unlabeled target data, we aim t o find representatives in the target data, which can not only represent and asso ciate data points belonging to each labeled category, but also discover novel ca tegories in the target data, if any. To leverage labeled source data, we guide r epresentative selection from labeled source to unlabeled target. We propose a jo

int optimization framework which alternately optimizes (1) representative selection in the target data and (2) discriminative feature learning from both the source and the target for better representative selection. Experiments on image and video datasets demonstrate that our proposed approach not only finds better representatives, but also can discover novel categories in the target data that are not in the source.
****

The Domain Transform Solver
Akash Bapat, Jan-Michael Frahm; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6014-6023
We present a novel framework for edge-aware optimization that is an order of magnitude faster than the state of the art while maintaining comparable results. Our key insight is that the optimization can be formulated by leveraging properties of the domain transform, a method for edge-aware filtering that defines a distance-preserving 1D mapping of the input space. This enables our method to improve performance for a wide variety of problems including stereo, depth super-resolution, render from defocus, colorization, and especially high-resolution depth filtering, while keeping the computational complexity linear in the number of pixels. Our method is highly parallelizable and adaptable, and it has demonstrable linear scalability with respect to image resolutions. We provide a comprehensive evaluation of our method w.r.t speed and accuracy for a variety of tasks.
****

CapSal: Leveraging Captioning to Boost Semantics for Salient Object Detection
Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, You He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6024-6033
Detecting salient objects in cluttered scenes is a big challenge. To address this problem, we argue that the model needs to learn discriminative semantic features for salient objects. To this end, we propose to leverage captioning as an auxiliary semantic task to boost salient object detection in complex scenarios. Specifically, we develop a CapSal model which consists of two sub-networks, the Image Captioning Network (ICN) and the Local-Global Perception Network (LGPN). ICN encodes the embedding of a generated caption to capture the semantic information of major objects in the scene, while LGPN incorporates the captioning embedding with local-global visual contexts for predicting the saliency map. ICN and LGPN are jointly trained to model high-level semantics as well as visual saliency. Extensive experiments demonstrate the effectiveness of image captioning in boosting the performance of salient object detection. In particular, our model performs significantly better than the state-of-the-art methods on several challenging datasets of complex scenarios.
****

Phase-Only Image Based Kernel Estimation for Single Image Blind Deblurring
Liyuan Pan, Richard Hartley, Miaomiao Liu, Yuchao Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6034-6043
The image motion blurring process is generally modelled as the convolution of a blur kernel with a latent image. Therefore, the estimation of the blur kernel is essentially important for blind image deblurring. Unlike existing approaches which focus on approaching the problem by enforcing various priors on the blur kernel and the latent image, we are aiming at obtaining a high quality blur kernel directly by studying the problem in the frequency domain. We show that the auto-correlation of the absolute phase-only image 1 can provide faithful information about the motion (e.g., the motion direction and magnitude, we call it the motion pattern in this paper.) that caused the blur, leading to a new and efficient blur kernel estimation approach. The blur kernel is then refined and the sharp image is estimated by solving an optimization problem by enforcing a regularization on the blur kernel and the latent image. We further extend our approach to handle non-uniform blur, which involves spatially varying blur kernels. Our approach is evaluated extensively on synthetic and real data and shows good results compared to the state-of-the-art deblurring approaches.

****
Hierarchical Discrete Distribution Decomposition for Match Density Estimation
Zhichao Yin, Trevor Darrell, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6044-6053
Explicit representations of the global match distributions of pixel-wise correspondences between pairs of images are desirable for uncertainty estimation and downstream applications. However, the computation of the match density for each pixel may be prohibitively expensive due to the large number of candidates. In this paper, we propose Hierarchical Discrete Distribution Decomposition (HD^3), a framework suitable for learning probabilistic pixel correspondences in both optical flow and stereo matching. We decompose the full match density into multiple scales hierarchically, and estimate the local matching distributions at each scale conditioned on the matching and warping at coarser scales. The local distributions can then be composed together to form the global match density. Despite its simplicity, our probabilistic method achieves state-of-the-art results for both optical flow and stereo matching on established benchmarks. We also find the estimated uncertainty is a good indication of the reliability of the predicted correspondences.
****
FOCNet: A Fractional Optimal Control Network for Image Denoising
Xixi Jia, Sanyang Liu, Xiangchu Feng, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6054-6063
Deep convolutional neural networks (DCNN) have been successfully used in many low-level vision problems such as image denoising. Recent studies on the mathematical foundation of DCNN has revealed that the forward propagation of DCNN corresponds to a dynamic system, which can be described by an ordinary differential equation (ODE) and solved by the optimal control method. However, most of these methods employ integer-order differential equation, which has local connectivity in time space and cannot describe the long-term memory of the system. Inspired by the fact that the fractional-order differential equation has long-term memory, in this paper we develop an advanced image denoising network, namely FOCNet, by solving a fractional optimal control (FOC) problem. Specifically, the network structure is designed based on the discretization of a fractional-order differential equation, which enjoys long-term memory in both forward and backward passes. Besides, multi-scale feature interactions are introduced into the FOCNet to strengthen the control of the dynamic system. Extensive experiments demonstrate the leading performance of the proposed FOCNet on image denoising. Code will be made available.
****
Orthogonal Decomposition Network for Pixel-Wise Binary Classification
Chang Liu, Fang Wan, Wei Ke, Zhuowei Xiao, Yuan Yao, Xiaosong Zhang, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6064-6073
The weight sharing scheme and spatial pooling operations in Convolutional Neural Networks (CNNs) introduce semantic correlation to neighboring pixels on feature maps and therefore deteriorate their pixel-wise classification performance. In this paper, we implement an Orthogonal Decomposition Unit (ODU) that transforms a convolutional feature map into orthogonal bases targeting at de-correlating neighboring pixels on convolutional features. In theory, complete orthogonal decomposition produces orthogonal bases which can perfectly reconstruct any binary mask (ground-truth). In practice, we further design incomplete orthogonal decomposition focusing on de-correlating local patches which balances the reconstruction performance and computational cost. Fully Convolutional Networks (FCNs) implemented with ODUs, referred to as Orthogonal Decomposition Networks (ODNs), learn de-correlated and complementary convolutional features and fuse such features in a pixel-wise selective manner. Over pixel-wise binary classification tasks for two-dimensional image processing, specifically skeleton detection, edge detection, and saliency detection, and one-dimensional keypoint detection, specifically S-wave arrival time detection for earthquake localization, ODNs consistently imp

roves the state-of-the-arts with significant margins.
****

Multi-Source Weak Supervision for Saliency Detection
Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6074-6083
The high cost of pixel-level annotations makes it appealing to train saliency detection models with weak supervision. However, a single weak supervision source usually does not contain enough information to train a well-performing model. To this end, we propose a unified framework to train saliency detection models with diverse weak supervision sources. In this paper, we use category labels, captions, and unlabelled data for training, yet other supervision sources can also be plugged into this flexible framework. We design a classification network (CNet) and a caption generation network (PNet), which learn to predict object categories and generate captions, respectively, meanwhile highlight the most important regions for corresponding tasks. An attention transfer loss is designed to transmit supervision signal between networks, such that the network designed to be trained with one supervision source can benefit from another. An attention coherence loss is defined on unlabelled data to encourage the networks to detect generally salient regions instead of task-specific regions. We use CNet and PNet to generate pixel-level pseudo labels to train a saliency prediction network (SNet). During the testing phases, we only need SNet to predict saliency maps. Experiments demonstrate the performance of our method compares favourably against unsupervised and weakly supervised methods and even some supervised methods.
****

ComDefend: An Efficient Image Compression Model to Defend Adversarial Examples
Xiaojun Jia, Xingxing Wei, Xiaochun Cao, Hassan Foroosh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6084-6092
Deep neural networks (DNNs) have been demonstrated to be vulnerable to adversarial examples. Specifically, adding imperceptible perturbations to clean images can fool the well trained deep neural networks. In this paper, we propose an end-to-end image compression model to defend adversarial examples: ComDefend. The proposed model consists of a compression convolutional neural network (ComCNN) and a reconstruction convolutional neural network (ResCNN). The ComCNN is used to maintain the structure information of the original image and purify adversarial perturbations. And the ResCNN is used to reconstruct the original image with high quality. In other words, ComDefend can transform the adversarial image to its clean version, which is then fed to the trained classifier. Our method is a pre-processing module, and does not modify the classifier's structure during the whole process. Therefore it can be combined with other model-specific defense models to jointly improve the classifier's robustness. A series of experiments conducted on MNIST, CIFAR10 and ImageNet show that the proposed method outperforms the state-of-the-art defense methods, and is consistently effective to protect classifiers against adversarial attacks.
****

Combinatorial Persistency Criteria for Multicut and Max-Cut
Jan-Hendrik Lange, Bjoern Andres, Paul Swoboda; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6093-6102
In combinatorial optimization, partial variable assignments are called persistent if they agree with some optimal solution. We propose persistency criteria for the multicut and max-cut problem as well as fast combinatorial routines to verify them. The criteria that we derive are based on mappings that improve feasible multicuts, respectively cuts. Our elementary criteria can be checked enumeratively. The more advanced ones rely on fast algorithms for upper and lower bounds for the respective cut problems and max-flow techniques for auxiliary min-cut problems. Our methods can be used as a preprocessing technique for reducing problem sizes or for computing partial optimality guarantees for solutions output by heuristic solvers. We show the efficacy of our methods on instances of both problems from computer vision, biomedical image analysis and statistical physics.

****
S4Net: Single Stage Salient-Instance Segmentation
Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, Shi-Min Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6103-6112
We consider an interesting problem---salient instance segmentation. Other than producing approximate bounding boxes, our network also outputs high-quality instance-level segments. Taking into account the category-independent property of each target, we design a single stage salient instance segmentation framework, with a novel segmentation branch. Our new branch regards not only local context inside each detection window but also its surrounding context, enabling us to distinguish the instances in the same scope even with obstruction. Our network is end-to-end trainable and runs at a fast speed (40 fps when processing an image with resolution 320 x 320). We evaluate our approach on a public available benchmark and show that it outperforms other alternative solutions. We also provide a thorough analysis of the design choices to help readers better understand the functions of each part of our network. The source code can be found at https://github.com/RuochenFan/S4Net.
****
A Decomposition Algorithm for the Sparse Generalized Eigenvalue Problem
Ganzhao Yuan, Li Shen, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6113-6122
The sparse generalized eigenvalue problem arises in a number of standard and modern statistical learning models, including sparse principal component analysis, sparse Fisher discriminant analysis, and sparse canonical correlation analysis. However, this problem is difficult to solve since it is NP-hard. In this paper, we consider a new effective decomposition method to tackle this problem. Specifically, we use random or/and swapping strategies to find a working set and perform global combinatorial search over the small subset of variables. We consider a bisection search method and a coordinate descent method for solving the quadratic fractional programming subproblem. In addition, we provide some theoretical analysis for the proposed method. Our experiments on synthetic data and real-world data have shown that our method significantly and consistently outperforms existing solutions in term of accuracy.
****
Polynomial Representation for Persistence Diagram
Zhichao Wang, Qian Li, Gang Li, Guandong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6123-6132
Persistence diagram (PD) has been considered as a compact descriptor for topological data analysis (TDA). Unfortunately, PD cannot be directly used in machine learning methods since it is a multiset of points. Recent efforts have been devoted to transforming PDs into vectors to accommodate machine learning methods. However, they share one common shortcoming: the mapping of PDs to a feature representation depends on a pre-defined polynomial. To address this limitation, this paper proposes an algebraic representation for PDs, i.e., polynomial representation. In this work, we discover a set of general polynomials that vanish on vectorized PDs and extract the task-adapted feature representation from these polynomials. We also prove two attractive properties of the proposed polynomial representation, i.e., stability and linear separability. Experiments also show that our method compares favorably with state-of-the-art TDA methods.
****
Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks
Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6133-6142
Crowd counting has recently attracted increasing interest in computer vision but remains a challenging problem. In this paper, we propose a trellis encoder-decoder network (TEDnet) for crowd counting, which focuses on generating high-quality density estimation maps. The major contributions are four-fold. First, we develop a new trellis architecture that incorporates multiple decoding paths to hier

archically aggregate features at different encoding stages, which improves the representative capability of convolutional features for large variations in objects. Second, we employ dense skip connections interleaved across paths to facilitate sufficient multi-scale feature fusions, which also helps TEDnet to absorb the supervision information. Third, we propose a new combinatorial loss to enforce similarities in local coherence and spatial correlation between maps. By distributedly imposing this combinatorial loss on intermediate outputs, TEDnet can improve the back-propagation process and alleviate the gradient vanishing problem. Finally, on four widely-used benchmarks, our TEDnet achieves the best overall performance in terms of both density map quality and counting accuracy, with an improvement up to 14% in MAE metric. These results validate the effectiveness of TEDnet for crowd counting.

****

Cross-Atlas Convolution for Parameterization Invariant Learning on Textured Mesh Surface

Shiwei Li, Zixin Luo, Mingmin Zhen, Yao Yao, Tianwei Shen, Tian Fang, Long Quan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6143-6152

We present a convolutional network architecture for direct feature learning on mesh surfaces through their atlases of texture maps. The texture map encodes the parameterization from 3D to 2D domain, rendering not only RGB values but also rasterized geometric features if necessary. Since the parameterization of texture map is not pre-determined, and depends on the surface topologies, we therefore introduce a novel cross-atlas convolution to recover the original mesh geodesic neighborhood, so as to achieve the invariance property to arbitrary parameterization. The proposed module is integrated into classification and segmentation architectures, which takes the input texture map of a mesh, and infers the output predictions. Our method not only shows competitive performances on classification and segmentation public benchmarks, but also paves the way for the broad mesh surfaces learning.

****

Deep Surface Normal Estimation With Hierarchical RGB-D Fusion

Jin Zeng, Yanfeng Tong, Yunmu Huang, Qiong Yan, Wenxiu Sun, Jing Chen, Yongtian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6153-6162

The growing availability of commodity RGB-D cameras has boosted the applications in the field of scene understanding. However, as a fundamental scene understanding task, surface normal estimation from RGB-D data lacks thorough investigation. In this paper, a hierarchical fusion network with adaptive feature re-weighting is proposed for surface normal estimation from a single RGB-D image. Specifically, the features from color image and depth are successively integrated at multiple scales to ensure global surface smoothness while preserving visually salient details. Meanwhile, the depth features are re-weighted with a confidence map estimated from depth before merging into the color branch to avoid artifacts caused by input depth corruption. Additionally, a hybrid multi-scale loss function is designed to learn accurate normal estimation given noisy ground-truth dataset. Extensive experimental results validate the effectiveness of the fusion strategy and the loss design, outperforming state-of-the-art normal estimation schemes.

****

Knowledge-Embedded Routing Network for Scene Graph Generation

Tianshui Chen, Weihao Yu, Riquan Chen, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6163-6171

To understand a scene in depth not only involves locating/recognizing individual objects, but also requires to infer the relationships and interactions among them. However, since the distribution of real-world relationships is seriously unbalanced, existing methods perform quite poorly for the less frequent relationships. In this work, we find that the statistical correlations between object pairs and their relationships can effectively regularize semantic space and make prediction less ambiguous, and thus well address the unbalanced distribution issue.

To achieve this, we incorporate these statistical correlations into deep neural networks to facilitate scene graph generation by developing a Knowledge-Embedded Routing Network. More specifically, we show that the statistical correlations between objects appearing in images and their relationships, can be explicitly represented by a structured knowledge graph, and a routing mechanism is learned to propagate messages through the graph to explore their interactions. Extensive experiments on the large-scale Visual Genome dataset demonstrate the superiority of the proposed method over current state-of-the-art competitors.
****

An End-To-End Network for Panoptic Segmentation
Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, Wei Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6172-6181
Panoptic segmentation, which needs to assign a category label to each pixel and segment each object instance simultaneously, is a challenging topic. Traditionally, the existing approaches utilize two independent models without sharing features, which makes the pipeline inefficient to implement. In addition, a heuristic method is usually employed to merge the results. However, the overlapping relationship between object instances is difficult to determine without sufficient context information during the merging process. To address the problems, we propose a novel end-to-end Occlusion Aware Network (OANet) for panoptic segmentation, which can efficiently and effectively predict both the instance and stuff segmentation in a single network. Moreover, we introduce a novel spatial ranking module to deal with the occlusion problem between the predicted instances. Extensive experiments have been done to validate the performance of our proposed method and promising results have been achieved on the COCO Panoptic benchmark.
****

Fast and Flexible Indoor Scene Synthesis via Deep Convolutional Generative Models
Daniel Ritchie, Kai Wang, Yu-An Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6182-6190
We present a new, fast and flexible pipeline for indoor scene synthesis that is based on deep convolutional generative models. Our method operates on a top-down image-based representation, and inserts objects iteratively into the scene by predict their category, location, orientation and size with separate neural network modules. Our pipeline naturally supports automatic completion of partial scenes, as well as synthesis of complete scenes, without any modifications. Our method is significantly faster than the previous image-based method, and generates results that outperforms it and other state-of-the-art deep generative scene models in terms of faithfulness to training data and perceived visual quality.
****

Marginalized Latent Semantic Encoder for Zero-Shot Learning
Zhengming Ding, Hongfu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6191-6199
Zero-shot learning has been well explored to precisely identify new unobserved classes through a visual-semantic function obtained from the existing objects. However, there exist two challenging obstacles: one is that the human-annotated semantics are insufficient to fully describe the visual samples; the other is the domain shift across existing and new classes. In this paper, we attempt to exploit the intrinsic relationship in the semantic manifold when given semantics are not enough to describe the visual objects, and enhance the generalization ability of the visual-semantic function with marginalized strategy. Specifically, we design a Marginalized Latent Semantic Encoder (MLSE), which is learned on the augmented seen visual features and the latent semantic representation. Meanwhile, latent semantics are discovered under an adaptive graph reconstruction scheme based on the provided semantics. Consequently, our proposed algorithm could enrich visual characteristics from seen classes, and well generalize to unobserved classes. Experimental results on zero-shot benchmarks demonstrate that the proposed model delivers superior performance over the state-of-the-art zero-shot learning approaches.

****
Scale-Adaptive Neural Dense Features: Learning via Hierarchical Context Aggregation
Jaime Spencer, Richard Bowden, Simon Hadfield; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6200-6209
How do computers and intelligent agents view the world around them? Feature extraction and representation constitutes one the basic building blocks towards answering this question. Traditionally, this has been done with carefully engineered hand-crafted techniques such as HOG, SIFT or ORB. However, there is no "one size fits all" approach that satisfies all requirements. In recent years, the rising popularity of deep learning has resulted in a myriad of end-to-end solutions to many computer vision problems. These approaches, while successful, tend to lack scalability and can't easily exploit information learned by other systems. Instead, we propose SAND features, a dedicated deep learning solution to feature extraction capable of providing hierarchical context information. This is achieved by employing sparse relative labels indicating relationships or similarity of dissimilarity between image locations. The nature of these labels results in an almost infinite set of dissimilar examples to choose from. We demonstrate how the selection of negative examples during training can be used to modify the feature space and vary it's learned properties. To demonstrate the generality of this approach, we apply the proposed features to a multitude of tasks, each requiring different properties. This includes disparity estimation, semantic segmentation, self-localisation and SLAM. In all cases, we show how incorporating SAND features results in better or comparable results to the baseline, whilst requiring little to no additional training.
****
Unsupervised Embedding Learning via Invariant and Spreading Instance Feature
Mang Ye, Xu Zhang, Pong C. Yuen, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6210-6219
This paper studies the unsupervised embedding learning problem, which requires an effective similarity measurement between samples in low-dimensional embedding space. Motivated by the positive concentrated and negative separated properties observed from category-wise supervised learning, we propose to utilize the instance-wise supervision to approximate these properties, which aims at learning data augmentation invariant and instance spread-out features. To achieve this goal, we propose a novel instance based softmax embedding method, which directly optimizes the `real' instance features on top of the softmax function. It achieves significantly faster learning speed and higher accuracy than all existing methods. The proposed method performs well for both seen and unseen testing categories with cosine similarity. It also achieves competitive performance even without pre-trained network over samples from fine-grained categories.
****
AOGNets: Compositional Grammatical Architectures for Deep Learning
Xilai Li, Xi Song, Tianfu Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6220-6230
Neural architectures are the foundation for improving performance of deep neural networks (DNNs). This paper presents deep compositional grammatical architectures which harness the best of two worlds: grammar models and DNNs. The proposed architectures integrate compositionality and reconfigurability of the former and the capability of learning rich features of the latter in a principled way. We utilize AND-OR Grammar (AOG) as network generator in this paper and call the resulting networks AOGNets. An AOGNet consists of a number of stages each of which is composed of a number of AOG building blocks. An AOG building block splits its input feature map into N groups along feature channels and then treat it as a sentence of N words. It then jointly realizes a phrase structure grammar and a dependency grammar in bottom-up parsing the "sentence" for better feature exploration and reuse. It provides a unified framework for the best practices developed in state-of-the-art DNNs. In experiments, AOGNet is tested in the ImageNet-1K classification benchmark and the MS-COCO object detection and segmentation benchmark. In ImageNet-1K, AOGNet obtains better performance than ResNet and most of its

variants, ResNeXt and its attention based variants such as SENet, DenseNet and DualPathNet. AOGNet also obtains the best model interpretability score using network dissection. AOGNet further shows better potential in adversarial defense. In MS-COCO, AOGNet obtains better performance than the ResNet and ResNeXt backbones in Mask R-CNN.

****

A Robust Local Spectral Descriptor for Matching Non-Rigid Shapes With Incompatible Shape Structures

Yiqun Wang, Jianwei Guo, Dong-Ming Yan, Kai Wang, Xiaopeng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6231-6240

Constructing a robust and discriminative local descriptor for 3D shape is a key component of many computer vision applications. Although existing learning-based approaches can achieve good performance in some specific benchmarks, they usually fail to learn enough information from shapes with different shape types and structures (e.g., spatial resolution, connectivity, transformations, etc.) Focusing on this issue, in this paper, we present a more discriminative local descriptor for deformable 3D shapes with incompatible structures. Based on the spectral embedding using the Laplace-Beltrami framework on the surface, we first construct a novel local spectral feature which shows great resilience to change in mesh resolution, triangulation, transformation. Then the multi-scale local spectral features around each vertex are encoded into a `geometry image', called vertex spectral image, in a very compact way. Such vertex spectral images can be efficiently trained to learn local descriptors using a triplet neural network. Finally, for training and evaluation, we present a new benchmark dataset by extending the widely used FAUST dataset. We utilize a remeshing approach to generate modified shapes with different structures. We evaluate the proposed approach thoroughly and make an extensive comparison to demonstrate that our approach outperforms recent state-of-the-art methods on this benchmark.

****

Context and Attribute Grounded Dense Captioning

Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6241-6250

Dense captioning aims at simultaneously localizing semantic regions and describing these regions-of-interest (ROIs) with short phrases or sentences in natural language. Previous studies have shown remarkable progresses, but they are often vulnerable to the aperture problem that a caption generated by the features inside one ROI lacks contextual coherence with its surrounding context in the input image. In this work, we investigate contextual reasoning based on multi-scale message propagations from the neighboring contents to the target ROIs. To this end, we design a novel end-to-end context and attribute grounded dense captioning framework consisting of 1) a contextual visual mining module and 2) a multi-level attribute grounded description generation module. Knowing that captions often co-occur with the linguistic attributes (such as who, what and where), we also incorporate an auxiliary supervision from hierarchical linguistic attributes to augment the distinctiveness of the learned captions. Extensive experiments and ablation studies on Visual Genome dataset demonstrate the superiority of the proposed model in comparison to state-of-the-art methods.

****

Spot and Learn: A Maximum-Entropy Patch Sampler for Few-Shot Image Classification

Wen-Hsuan Chu, Yu-Jhe Li, Jing-Cheng Chang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6251-6260

Few-shot learning (FSL) requires one to learn from object categories with a small amount of training data (as novel classes), while the remaining categories (as base classes) contain a sufficient amount of data for training. It is often desirable to transfer knowledge from the base classes and derive dominant features efficiently for the novel samples. In this work, we propose a sampling method th

at de-correlates an image based on maximum entropy reinforcement learning, and e
xtracts varying sequences of patches on every forward-pass with discriminative i
nformation observed. This can be viewed as a form of "learned" data augmentation
 in the sense that we search for different sequences of patches within an image
and performs classification with aggregation of the extracted features, resultin
g in improved FSL performances. In addition, our positive and negative sampling
 policies along with a newly defined reward function would favorably improve the
 effectiveness of our model. Our experiments on two benchmark datasets confirm th
e effectiveness of our framework and its superiority over recent FSL approaches.
****

Interpreting CNNs via Decision Trees

Quanshi Zhang, Yu Yang, Haotian Ma, Ying Nian Wu; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6261-62
70

This paper aims to quantitatively explain the rationales of each prediction that
 is made by a pre-trained convolutional neural network (CNN). We propose to lear
n a decision tree, which clarifies the specific reason for each prediction made
by the CNN at the semantic level. I.e., the decision tree decomposes feature rep
resentations in high conv-layers of the CNN into elementary concepts of object p
arts. In this way, the decision tree tells people which object parts activate wh
ich filters for the prediction and how much each object part contributes to the
prediction score. Such semantic and quantitative explanations for CNN prediction
s have specific values beyond the traditional pixel-level analysis of CNNs. More
 specifically, our method mines all potential decision modes of the CNN, where e
ach mode represents a typical case of how the CNN uses object parts for predicti
on. The decision tree organizes all potential decision modes in a coarse-to-fine
 manner to explain CNN predictions at different fine-grained levels. Experiments
 have demonstrated the effectiveness of the proposed method.
****

Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Capti
oning

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, In So Kweon; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6271
-6280

Our goal in this work is to train an image captioning model that generates more
dense and informative captions. We introduce "relational captioning," a novel im
age captioning task which aims to generate multiple captions with respect to rel
ational information between objects in an image. Relational captioning is a fram
ework that is advantageous in both diversity and amount of information, leading
to image understanding based on relationships. Part-of-speech (POS, i.e. subject
-object-predicate categories) tags can be assigned to every English word. We lev
erage the POS as a prior to guide the correct sequence of words in a caption. To
 this end, we propose a multi-task triple-stream network (MTTSNet) which consist
s of three recurrent units for the respective POS and jointly performs POS predi
ction and captioning. We demonstrate more diverse and richer representations gen
erated by the proposed model against several baselines and competing methods.
****

Deep Modular Co-Attention Networks for Visual Question Answering

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, Qi Tian; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6281-6
290

Visual Question Answering (VQA) requires a fine-grained and simultaneous underst
anding of both the visual content of images and the textual content of questions
. Therefore, designing an effective `co-attention' model to associate key words
in questions with key objects in images is central to VQA performance. So far, m
ost successful attempts at co-attention learning have been achieved by using sha
llow models, and deep co-attention models show little improvement over their sha
llow counterparts. In this paper, we propose a deep Modular Co-Attention Network
 (MCAN) that consists of Modular Co-Attention (MCA) layers cascaded in depth. Ea
ch MCA layer models the self-attention of questions and images, as well as the q

uestion-guided-attention of images jointly using a modular composition of two ba
sic attention units. We quantitatively and qualitatively evaluate MCAN on the be
nchmark VQA-v2 dataset and conduct extensive ablation studies to explore the rea
sons behind MCAN's effectiveness. Experimental results demonstrate that MCAN sig
nificantly outperforms the previous state-of-the-art. Our best single model deli
vers 70.63% overall accuracy on the test-dev set.
****

Synthesizing Environment-Aware Activities via Activity Sketches
Yuan-Hong Liao, Xavier Puig, Marko Boben, Antonio Torralba, Sanja Fidler; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2019, pp. 6291-6299
In order to learn to perform activities from demonstrations or descriptions, age
nts need to distill what the essence of the given activity is, and how it can be
adapted to new environments. In this work, we address the problem: environment-
aware program generation. Given a visual demonstration or a description of an ac
tivity, we generate program sketches representing the essential instructions and
propose a model to flesh these into full programs representing the actions need
ed to perform the activity under the presented environmental constraints. To thi
s end, we build upon VirtualHome, to create a new dataset VirtualHome-Env, where
we collect program sketches to represent activities and match programs with env
ironments that can afford them. Furthermore, we construct a knowledge base to sa
mple realistic environments and another knowledge base to seek out the programs
under the sampled environments. Finally, we propose RNN-ResActGraph, a network
that generates a program from a given sketch and an environment graph and tracks
the changes in the environment induced by the program.
****

Self-Critical N-Step Training for Image Captioning
Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, Wen Gao; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 6300-6308
Existing methods for image captioning are usually trained by cross entropy loss,
which leads to exposure bias and the inconsistency between the optimizing funct
ion and evaluation metrics. Recently it has been shown that these two issues can
be addressed by incorporating techniques from reinforcement learning, where one
of the popular techniques is the advantage actor-critic algorithm that calculat
es per-token advantage by estimating state value with a parametrized estimator a
t the cost of introducing estimation bias. In this paper, we estimate state valu
e without using a parametrized value estimator. With the properties of image cap
tioning, namely, the deterministic state transition function and the sparse rewa
rd, state value is equivalent to its preceding state-action value, and we reform
ulate advantage function by simply replacing the former with the latter. Moreove
r, the reformulated advantage is extended to n-step, which can generally increas
e the absolute value of the mean of reformulated advantage while lowering varian
ce. Then two kinds of rollout are adopted to estimate state-action value, which
we call self-critical n-step training. Empirically we find that our method can o
btain better performance compared to the state-of-the-art methods that use the s
equence level advantage and parametrized estimator respectively on the widely us
ed MSCOCO benchmark.
****

Multi-Target Embodied Question Answering
Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, Dh
ruv Batra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2019, pp. 6309-6318
Embodied Question Answering (EQA) is a relatively new task where an agent is ask
ed to answer questions about its environment from egocentric perception. EQA as
introduced in [8] makes the fundamental assumption that every question, e.g., "
what color is the car?", has exactly one target ("car") being inquired about. Th
is assumption puts a direct limitation on the abilities of the agent. We presen
t a generalization of EQA -- Multi-Target EQA (MT-EQA). Specifically, we study
questions that have multiple targets in them, such as "Is the dresser in the bed

room bigger than the oven in the kitchen?", where the agent has to navigate to multiple locations ("dresser in bedroom", "oven in kitchen") and perform comparative reasoning ("dresser" bigger than "oven") before it can answer a question. Such questions require the development of entirely new modules or components in the agent. To address this, we propose a modular architecture composed of a program generator, a controller, a navigator, and a VQA module. The program generator converts the given question into sequential executable sub-programs; the navigator guides the agent to multiple locations pertinent to the navigation-related sub-programs; and the controller learns to select relevant observations along its path. These observations are then fed to the VQA module to predict the answer. We perform detailed analysis for each of the model components and show that our joint model can outperform previous methods and strong baselines by a significant margin.

****

Visual Question Answering as Reading Comprehension

Hui Li, Peng Wang, Chunhua Shen, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6319-6328

Visual question answering (VQA) demands simultaneous comprehension of both the image visual content and natural language questions. In some cases, the reasoning needs the help of common sense or general knowledge which usually appear in the form of text. Current methods jointly embed both the visual information and the textual feature into the same space. Nevertheless, how to model the complex interactions between the two different modalities is not an easy work. In contrast to struggling on multimodal feature fusion, in this paper, we propose to unify all the input information by natural language so as to convert VQA into a machine reading comprehension problem. With this transformation, our method not only can tackle VQA datasets that focus on observation based questions, but can also be naturally extended to handle knowledge-based VQA which requires to explore large-scale external knowledge base. It is a step towards being able to exploit large volumes of text and natural language processing techniques to address VQA problem. Two types of models are proposed to deal with open-ended VQA and multiple-choice VQA respectively. We evaluate our models on three VQA benchmarks. The comparable performance with the state-of-the-art demonstrates the effectiveness of the proposed method.

****

StoryGAN: A Sequential Conditional GAN for Story Visualization

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6329-6338

In this work, we propose a new task called Story Visualization. Given a multi-sentence paragraph, the story is visualized by generating a sequence of images, one for each sentence. In contrast to video generation, story visualization focuses less on the continuity in generated images (frames), but more on the global consistency across dynamic scenes and characters -- a challenge that has not been addressed by any single-image or video generation methods. Therefore, we propose a new story-to-image-sequence generation model, StoryGAN, based on the sequential conditional GAN framework. Our model is unique in that it consists of a deep Context Encoder that dynamically tracks the story flow, and two discriminators at the story and image levels, to enhance the image quality and the consistency of the generated sequences. To evaluate the model, we modified existing datasets to create the CLEVR-SV and Pororo-SV datasets. Empirically, StoryGAN outperformed state-of-the-art models in image quality, contextual consistency metrics, and human evaluation.

****

Noise-Aware Unsupervised Deep Lidar-Stereo Fusion

Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6339-6348

In this paper, we present LidarStereoNet, the first unsupervised Lidar-stereo fu

sion network, which can be trained in an end-to-end manner without the need of g round truth depth maps. By introducing a novel "Feedback Loop" to connect the ne twork input with output, LidarStereoNet could tackle both noisy Lidar points and misalignment between sensors that have been ignored in existing Lidar-stereo fu sion work. Besides, we propose to incorporate the piecewise planar model into th e network learning to further constrain depths to conform to the underlying 3D g eometry. Extensive quantitative and qualitative evaluations on both real and syn thetic datasets demonstrate the superiority of our method, which outperforms sta te-of-the-art stereo matching, depth completion and Lidar-Stereo fusion approach es significantly.

****

Versatile Multiple Choice Learning and Its Application to Vision Computing
Kai Tian, Yi Xu, Shuigeng Zhou, Jihong Guan; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6349-6357
Most existing ensemble methods aim to train the underlying embedded models indep endently and simply aggregate their final outputs via averaging or weighted voti ng. As many prediction tasks contain uncertainty, most of these ensemble methods just reduce variance of the predictions without considering the collaborations among the ensembles. Different from these ensemble methods, multiple choice lear ning (MCL) methods exploit the cooperation among all the embedded models to gene rate multiple diverse hypotheses. In this paper, a new MCL method, called vMCL (the abbreviation of versatile Multiple Choice Learning), is developed to extend the application scenarios of MCL methods by ensembling deep neural networks. Our vMCL method keeps the advantage of existing MCL methods while overcoming their major drawback, thus achieves better performance. The novelty of our vMCL lies i n three aspects: (1) a choice network is designed to learn the confidence level of each specialist which can provide the best prediction base on multiple hypoth eses; (2) a hinge loss is introduced to alleviate the overconfidence issue in MC L settings; (3) Easy to be implemented and can be trained in an end-to-end manne r, which is a very attractive feature for many real-world applications. Experime nts on image classification and image segmentation task show that vMCL outperfor ms the existing state-of-the-art MCL methods.

****

EV-Gait: Event-Based Robust Gait Recognition Using Dynamic Vision Sensors
Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, Hongkai Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2019, pp. 6358-6367
In this paper, we introduce a new type of sensing modality, the Dynamic Vision S ensors (Event Cameras), for the task of gait recognition. Compared with the trad itional RGB sensors, the event cameras have many unique advantages such as ultra low resources consumption, high temporal resolution and much larger dynamic ran ge. However, those cameras only produce noisy and asynchronous events of intensi ty changes rather than frames, where conventional vision-based gait recognition algorithms can't be directly applied. To address this, we propose a new Event-ba sed Gait Recognition (EV-Gait) approach, which exploits motion consistency to ef fectively remove noise, and uses a deep neural network to recognise gait from th e event streams. To evaluate the performance of EV-Gait, we collect two event-ba sed gait datasets, one from real-world experiments and the other by converting t he publicly available RGB gait recognition benchmark CASIA-B. Extensive experime nts show that EV-Gait can get nearly 96% recognition accuracy in the real-world settings, while on the CASIA-B benchmark it achieves comparable performance with state-of-the-art RGB-based gait recognition approaches.

****

ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Bea m CT Images
Zhiming Cui, Changjian Li, Wenping Wang; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6368-6377
This paper proposes a method that uses deep convolutional neural networks to ach ieve automatic and accurate tooth instance segmentation and identification from CBCT (cone beam CT) images for digital dentistry. The core of our method is a tw

o-stage network. In the first stage, an edge map is extracted from the input CBC T image to enhance image contrast along shape boundaries. Then this edge map and the input images are passed to the second stage. In the second stage, we build our network upon the 3D region proposal network (RPN) with a novel learned-simil arity matrix to help efficiently remove redundant proposals, speed up training a nd save GPU memory. To resolve the ambiguity in the identification task, we enco de teeth spatial relationships as an additional feature input in the identificat ion task, which helps to remarkably improve the identification accuracy. Our eva luation, comparison and comprehensive ablation studies demonstrate that our meth od produces accurate instance segmentation and identification results automatica lly and outperforms the state-of-the-art approaches. To the best of our knowledg e, our method is the first to use neural networks to achieve automatic tooth seg mentation and identification from CBCT images.

****

Modularized Textual Grounding for Counterfactual Resilience

Zhiyuan Fang, Shu Kong, Charless Fowlkes, Yezhou Yang; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 63 78-6388

Computer Vision applications often require a textual grounding module with preci sion, interpretability, and resilience to counterfactual inputs/queries. To achi eve high grounding precision, current textual grounding methods heavily rely on large-scale training data with manual annotations at the pixel level. Such annot ations are expensive to obtain and thus severely narrow the model's scope of rea l-world applications. Moreover, most of these methods sacrifice interpretability , generalizability, and they neglect the importance of being resilient to counte rfactual inputs. To address these issues, we propose a visual grounding system w hich is 1) end-to-end trainable in a weakly supervised fashion with only image-l evel annotations, and 2) counterfactually resilient owing to the modular design. Specifically, we decompose textual descriptions into three levels: entity, sema ntic attribute, color information, and perform compositional grounding progress ively. We validate our model through a series of experiments and demonstrate its improvement over the state-of-the-art methods. In particular, our model's perfo rmance not only surpasses other weakly/un-supervised methods and even approaches the strongly supervised ones, but also is interpretable for decision making and performs much better in face of counterfactual classes than all the others.

****

L3-Net: Towards Learning Based LiDAR Localization for Autonomous Driving

Weixin Lu, Yao Zhou, Guowei Wan, Shenhua Hou, Shiyu Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp . 6389-6398

We present L3-Net - a novel learning-based LiDAR localization system that achiev es centimeter-level localization accuracy, comparable to prior state-of-the-art systems with hand-crafted pipelines. Rather than relying on these hand-crafted m odules, we innovatively implement the use of various deep neural network structu res to establish a learning-based approach. L3-Net learns local descriptors spec ifically optimized for matching in different real-world driving scenarios. 3D co nvolutions over a cost volume built in the solution space significantly boosts t he localization accuracy. RNNs are demonstrated to be effective in modeling the vehicle's dynamics, yielding better temporal smoothness and accuracy. We compreh ensively validate the effectiveness of our approach using freshly collected data sets. Multiple trials of repetitive data collection over the same road and areas make our dataset ideal for testing localization systems. The SunnyvaleBigLoop sequences, with a year's time interval between the collected mapping and testing data, made it quite challenging, but the low localization error of our method i n these datasets demonstrates its maturity for real industrial implementation.

****

Panoptic Feature Pyramid Networks

Alexander Kirillov, Ross Girshick, Kaiming He, Piotr Dollar; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6399-6408

The recently introduced panoptic segmentation task has renewed our community's i nterest in unifying the tasks of instance segmentation (for thing classes) and s emantic segmentation (for stuff classes). However, current state-of-the-art meth ods for this joint task use separate and dissimilar networks for instance and se mantic segmentation, without performing any shared computation. In this work, we aim to unify these methods at the architectural level, designing a single netwo rk for both tasks. Our approach is to endow Mask R-CNN, a popular instance segme ntation method, with a semantic segmentation branch using a shared Feature Pyram id Network (FPN) backbone. Surprisingly, this simple baseline not only remains e ffective for instance segmentation, but also yields a lightweight, top-performin g method for semantic segmentation. In this work, we perform a detailed study of this minimally extended version of Mask R-CNN with FPN, which we refer to as Pa noptic FPN, and show it is a robust and accurate baseline for both tasks. Given its effectiveness and conceptual simplicity, we hope our method can serve as a s trong baseline and aid future research in panoptic segmentation.

****

Mask Scoring R-CNN

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, Xinggang Wang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2019, pp. 6409-6418

Letting a deep network be aware of the quality of its own predictions is an inte resting yet important problem. In the task of instance segmentation, the confide nce of instance classification is used as mask quality score in most instance se gmentation frameworks. However, the mask quality, quantified as the IoU between the instance mask and its ground truth, is usually not well correlated with clas sification score. In this paper, we study this problem and propose Mask Scoring R-CNN which contains a network block to learn the quality of the predicted insta nce masks. The proposed network block takes the instance feature and the corresp onding predicted mask together to regress the mask IoU. The mask scoring strateg y calibrates the misalignment between mask quality and mask score, and improves instance segmentation performance by prioritizing more accurate mask predictions during COCO AP evaluation. By extensive evaluations on the COCO dataset, Mask S coring R-CNN brings consistent and noticeable gain with different models and out performs the state-of-the-art Mask R-CNN. We hope our simple and effective appro ach will provide a new direction for improving instance segmentation. The source code of our method is available at https://github.com/zjhuang22/maskscoring_rcn n.

****

Reasoning-RCNN: Unifying Adaptive Global Reasoning Into Large-Scale Object Detec tion

Hang Xu, Chenhan Jiang, Xiaodan Liang, Liang Lin, Zhenguo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 , pp. 6419-6428

In this paper, we address the large-scale object detection problem with thousand s of categories, which poses severe challenges due to long-tail data distributio ns, heavy occlusions, and class ambiguities. However, the dominant object detect ion paradigm is limited by treating each object region separately without consid ering crucial semantic dependencies among objects. In this work, we introduce a novel Reasoning-RCNN to endow any detection networks the capability of adaptive global reasoning over all object regions by exploiting diverse human commonsense knowledge. Instead of only propagating the visual features on the image directl y, we evolve the high-level semantic representations of all categories globally to avoid distracted or poor visual features in the image. Specifically, built on feature representations of basic detection network, the proposed network first generates a global semantic pool by collecting the weights of previous classific ation layer for each category, and then adaptively enhances each object features via attending different semantic contexts in the global semantic pool. Rather t han propagating information from all semantic information that may be noisy, our adaptive global reasoning automatically discovers most relative categories for feature evolving. Our Reasoning-RCNN is light-weight and flexible enough to enha

nce any detection backbone networks, and extensible for integrating any knowledg e resources. Solid experiments on object detection benchmarks show the superiori ty of our Reasoning-RCNN, e.g. achieving around 16% improvement on VisualGenome, 37% on ADE in terms of mAP and 15% improvement on COCO.
****

Cross-Modality Personalization for Retrieval
Nils Murrugarra-Llerena, Adriana Kovashka; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6429-6438
Existing captioning and gaze prediction approaches do not consider the multiple facets of personality that affect how a viewer extracts meaning from an image. W hile there are methods that consider personalized captioning, they do not consid er personalized perception across modalities, i.e. how a person's way of looking at an image (gaze) affects the way they describe it (captioning). In this work, we propose a model for modeling cross-modality personalized retrieval. In addit ion to modeling gaze and captions, we also explicitly model the personality of t he users providing these samples. We incorporate constraints that encourage gaze and caption samples on the same image to be close in a learned space; we refer to this as content modeling. We also model style: we encourage samples provided by the same user to be close in a separate embedding space, regardless of the im age on which they were provided. To leverage the complementary information that content and style constraints provide, we combine the embeddings from both netwo rks. We show that our combined embeddings achieve better performance than existi ng approaches for cross-modal retrieval.
****

Composing Text and Image for Image Retrieval - an Empirical Odyssey
Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, James Hay s; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2019, pp. 6439-6448
In this paper, we study the task of image retrieval, where the input query is sp ecified in the form of an image plus some text that describes desired modificati ons to the input image. For example, we may present an image of the Eiffel tower , and ask the system to find images which are visually similar, but are modified in small ways, such as being taken at nighttime instead of during the day. o ta ckle this task, we embed the query (reference image plus modification text) and the target (images). The encoding function of the image text query learns a repr esentation, such that the similarity with the target image representation is hig h iff it is a "positive match". We propose a new way to combine image and text t hrough residual connection, that is designed for this retrieval task. We show th is outperforms existing approaches on 3 different datasets, namely Fashion-200k, MIT-States and a new synthetic dataset we create based on CLEVR. We also show t hat our approach can be used to perform image classification with compositionall y novel labels, and we outperform previous methods on MIT-States on this task.
****

Arbitrary Shape Scene Text Detection With Adaptive Text Region Representation
Xiaobing Wang, Yingying Jiang, Zhenbo Luo, Cheng-Lin Liu, Hyunsoo Choi, Sun gjin Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6449-6458
Scene text detection attracts much attention in computer vision, because it can be widely used in many applications such as real-time text translation, automati c information entry, blind person assistance, robot sensing and so on. Though ma ny methods have been proposed for horizontal and oriented texts, detecting irreg ular shape texts such as curved texts is still a challenging problem. To solve t he problem, we propose a robust scene text detection method with adaptive text r egion representation. Given an input image, a text region proposal network is fi rst used for extracting text proposals. Then, these proposals are verified and r efined with a refinement network. Here, recurrent neural network based adaptive text region representation is proposed for text region refinement, where a pair of boundary points are predicted each time step until no new points are found. I n this way, text regions of arbitrary shapes are detected and represented with a daptive number of boundary points. This gives more accurate description of text

regions. Experimental results on five benchmarks, namely, CTW1500, TotalText, ICDAR2013, ICDAR2015 and MSRA-TD500, show that the proposed method achieves state-of-the-art in scene text detection.
****

Adaptive NMS: Refining Pedestrian Detection in a Crowd
Songtao Liu, Di Huang, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6459-6468
Pedestrian detection in a crowd is a very challenging issue. This paper addresses this problem by a novel Non-Maximum Suppression (NMS) algorithm to better refine the bounding boxes given by detectors. The contributions are threefold: (1) we propose adaptive-NMS, which applies a dynamic suppression threshold to an instance, according to the target density; (2) we design an efficient subnetwork to learn density scores, which can be conveniently embedded into both the single-stage and two-stage detectors; and (3) we achieve state of the art results on the CityPersons and CrowdHuman benchmarks.
****

Point in, Box Out: Beyond Counting Persons in Crowds
Yuting Liu, Miaojing Shi, Qijun Zhao, Xiaofang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6469-6478
Modern crowd counting methods usually employ deep neural networks (DNN) to estimate crowd counts via density regression. Despite their significant improvements, the regression-based methods are incapable of providing the detection of individuals in crowds. The detection-based methods, on the other hand, have not been largely explored in recent trends of crowd counting due to the needs for expensive bounding box annotations. In this work, we instead propose a new deep detection network with only point supervision required. It can simultaneously detect the size and location of human heads and count them in crowds. We first mine useful person size information from point-level annotations and initialize the pseudo ground truth bounding boxes. An online updating scheme is introduced to refine the pseudo ground truth during training; while a locally-constrained regression loss is designed to provide additional constraints on the size of the predicted boxes in a local neighborhood. In the end, we propose a curriculum learning strategy to train the network from images of relatively accurate and easy pseudo ground truth first. Extensive experiments are conducted in both detection and counting tasks on several standard benchmarks, e.g. ShanghaiTech, UCF_CC_50, WiderFace, and TRANCOS datasets, and the results show the superiority of our method over the state-of-the-art.
****

Locating Objects Without Bounding Boxes
Javier Ribera, David Guera, Yuhao Chen, Edward J. Delp; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6479-6489
Recent advances in convolutional neural networks (CNN) have achieved remarkable results in locating objects in images. In these networks, the training procedure usually requires providing bounding boxes or the maximum number of expected objects. In this paper, we address the task of estimating object locations without annotated bounding boxes which are typically hand-drawn and time consuming to label. We propose a loss function that can be used in any fully convolutional network (FCN) to estimate object locations. This loss function is a modification of the average Hausdorff distance between two unordered sets of points. The proposed method has no notion of bounding boxes, region proposals, or sliding windows. We evaluate our method with three datasets designed to locate people's heads, pupil centers and plant centers. We outperform state-of-the-art generic object detectors and methods fine-tuned for pupil tracking.
****

FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery
Krishna Kumar Singh, Utkarsh Ojha, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6490-6499

We propose FineGAN, a novel unsupervised GAN framework, which disentangles the background, object shape, and object appearance to hierarchically generate images of fine-grained object categories. To disentangle the factors without supervision, our key idea is to use information theory to associate each factor to a latent code, and to condition the relationships between the codes in a specific way to induce the desired hierarchy. Through extensive experiments, we show that FineGAN achieves the desired disentanglement to generate realistic and diverse images belonging to fine-grained classes of birds, dogs, and cars. Using FineGAN's automatically learned features, we also cluster real images as a first attempt at solving the novel problem of unsupervised fine-grained object category discovery. Our code/models/demo can be found at https://github.com/kkanshul/finegan

****

Mutual Learning of Complementary Networks via Residual Correction for Improving Semi-Supervised Classification

Si Wu, Jichang Li, Cheng Liu, Zhiwen Yu, Hau-San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6500-6509

Deep mutual learning jointly trains multiple essential networks having similar properties to improve semi-supervised classification. However, the commonly used consistency regularization between the outputs of the networks may not fully leverage the difference between them. In this paper, we explore how to capture the complementary information to enhance mutual learning. For this purpose, we propose a complementary correction network (CCN), built on top of the essential networks, to learn the mapping from the output of one essential network to the ground truth label, conditioned on the features learnt by another. To make the second essential network increasingly complementary to the first one, this network is supervised by the corrected predictions. As a result, minimizing the prediction divergence between the two complementary networks can lead to significant performance gains in semi-supervised learning. Our experimental results demonstrate that the proposed approach clearly improves mutual learning between essential networks, and achieves state-of-the-art results on multiple semi-supervised classification benchmarks. In particular, the test error rates are reduced from previous 21.23% and 14.65% to 12.05% and 10.37% on CIFAR-10 with 1000 and 2000 labels, respectively.

****

Sampling Techniques for Large-Scale Object Detection From Sparsely Annotated Objects

Yusuke Niitani, Takuya Akiba, Tommi Kerola, Toru Ogawa, Shotaro Sano, Shuji Suzuki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6510-6518

Efficient and reliable methods for training of object detectors are in higher demand than ever, and more and more data relevant to the field is becoming available. However, large datasets like Open Images Dataset v4 (OID) are sparsely annotated, and some measure must be taken in order to ensure the training of a reliable detector. In order to take the incompleteness of these datasets into account, one possibility is to use pretrained models to detect the presence of the unverified objects. However, the performance of such a strategy depends largely on the power of the pretrained model. In this study, we propose part-aware sampling, a method that uses human intuition for the hierarchical relation between objects. In terse terms, our method works by making assumptions like "a bounding box for a car should contain a bounding box for a tire". We demonstrate the power of our method on OID and compare the performance against a method based on a pretrained model. Our method also won the first and second place on the public and private test sets of the Google AI Open Images Competition 2018.

****

Curls & Whey: Boosting Black-Box Adversarial Attacks

Yucheng Shi, Siyu Wang, Yahong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6519-6527

Image classifiers based on deep neural networks suffer from harassment caused by adversarial examples. Two defects exist in black-box iterative attacks that gen

erate adversarial examples by incrementally adjusting the noise-adding direction for each step. On the one hand, existing iterative attacks add noises monotonically along the direction of gradient ascent, resulting in a lack of diversity and adaptability of the generated iterative trajectories. On the other hand, it is trivial to perform adversarial attack by adding excessive noises, but currently there is no refinement mechanism to squeeze redundant noises. In this work, we propose Curls & Whey black-box attack to fix the above two defects. During Curls iteration, by combining gradient ascent and descent, we `curl' up iterative trajectories to integrate more diversity and transferability into adversarial examples. Curls iteration also alleviates the diminishing marginal effect in existing iterative attacks. The Whey optimization further squeezes the `whey' of noises by exploiting the robustness of adversarial perturbation. Extensive experiments on Imagenet and Tiny-Imagenet demonstrate that our approach achieves impressive decrease on noise magnitude in l2 norm. Curls & Whey attack also shows promising transferability against ensemble models as well as adversarially trained models. In addition, we extend our attack to the targeted misclassification, effectively reducing the difficulty of targeted attacks under black-box condition.
****

Barrage of Random Transforms for Adversarially Robust Defense
Edward Raff, Jared Sylvester, Steven Forsyth, Mark McLean; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6528-6537
Defenses against adversarial examples, when using the ImageNet dataset, are historically easy to defeat. The common understanding is that a combination of simple image transformations and other various defenses are insufficient to provide the necessary protection when the obfuscated gradient is taken into account. In this paper, we explore the idea of stochastically combining a large number of individually weak defenses into a single barrage of randomized transformations to build a strong defense against adversarial attacks. We show that, even after accounting for obfuscated gradients, the Barrage of Random Transforms (BaRT) is a resilient defense against even the most difficult attacks, such as PGD. BaRT achieves up to a 24x improvement in accuracy compared to previous work, and has even extended effectiveness out to a previously untested maximum adversarial perturbation of e=32.
****

Aggregation Cross-Entropy for Sequence Recognition
Zecheng Xie, Yaoxiong Huang, Yuanzhi Zhu, Lianwen Jin, Yuliang Liu, Lele Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6538-6547
In this paper, we propose a novel method, aggregation cross-entropy (ACE), for sequence recognition from a brand new perspective. The ACE loss function exhibits competitive performance to CTC and the attention mechanism, with much quicker implementation (as it involves only four fundamental formulas), faster inference\back-propagation (approximately O(1) in parallel), less storage requirement (no parameter and negligible runtime memory), and convenient employment (by replacing CTC with ACE). Furthermore, the proposed ACE loss function exhibits two noteworthy properties: (1) it can be directly applied for 2D prediction by flattening the 2D prediction into 1D prediction as the input and (2) it requires only characters and their numbers in the sequence annotation for supervision, which allows it to advance beyond sequence recognition, e.g., counting problem. The code is publicly available at https://github.com/summerlvsong/Aggregation-Cross-Entropy.
****

LaSO: Label-Set Operations Networks for Multi-Label Few-Shot Learning
Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, Alex M. Bronstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6548-6557
Example synthesis is one of the leading methods to tackle the problem of few-shot learning, where only a small number of samples per class are available. However, current synthesis approaches only address the scenario of a single category label per image. In this work, we propose a novel technique for synthesizing samp

les with multiple labels for the (yet unhandled) multi-label few-shot classification scenario. We propose to combine pairs of given examples in feature space, so that the resulting synthesized feature vectors will correspond to examples whose label sets are obtained through certain set operations on the label sets of the corresponding input pairs. Thus, our method is capable of producing a sample containing the intersection, union or set-difference of labels present in two input samples. As we show, these set operations generalize to labels unseen during training. This enables performing augmentation on examples of novel categories, thus, facilitating multi-label few-shot classifier learning. We conduct numerous experiments showing promising results for the label-set manipulation capabilities of the proposed approach, both directly (using the classification and retrieval metrics), and in the context of performing data augmentation for multi-label few-shot learning. We propose a benchmark for this new and challenging task and show that our method compares favorably to all the common baselines.
****

Few-Shot Learning With Localization in Realistic Settings
Davis Wertheimer, Bharath Hariharan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6558-6567
Traditional recognition methods typically require large, artificially-balanced training classes, while few-shot learning methods are tested on artificially small ones. In contrast to both extremes, real world recognition problems exhibit heavy-tailed class distributions, with cluttered scenes and a mix of coarse and fine-grained class distinctions. We show that prior methods designed for few-shot learning do not work out of the box in these challenging conditions, based on a new "meta-iNat" benchmark. We introduce three parameter-free improvements: (a) better training procedures based on adapting cross-validation to meta-learning, (b) novel architectures that localize objects using limited bounding box annotations before classification, and (c) simple parameter-free expansions of the feature space based on bilinear pooling. Together, these improvements double the accuracy of state-of-the-art models on meta-iNat while generalizing to prior benchmarks, complex neural architectures, and settings with substantial domain shift.
****

AdaGraph: Unifying Predictive and Continuous Domain Adaptation Through Graphs
Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6568-6577
The ability to categorize is a cornerstone of visual intelligence, and a key functionality for artificial, autonomous visual machines. This problem will never be solved without algorithms able to adapt and generalize across visual domains. Within the context of domain adaptation and generalization, this paper focuses on the predictive domain adaptation scenario, namely the case where no target data are available and the system has to learn to generalize from annotated source images plus unlabeled samples with associated metadata from auxiliary domains. Our contribution is the first deep architecture that tackles predictive domain adaptation, able to leverage over the information brought by the auxiliary domains through a graph. Moreover, we present a simple yet effective strategy that allows us to take advantage of the incoming target data at test time, in a continuous domain adaptation scenario. Experiments on three benchmark databases support the value of our approach.
****

Grounded Video Description
Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, Marcus Rohrbach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6578-6587
Video description is one of the most challenging problems in vision and language understanding due to the large variability both on the video and language side. Models, hence, typically shortcut the difficulty in recognition and generate plausible sentences that are based on priors but are not necessarily grounded in the video. In this work, we explicitly link the sentence to the evidence in the video by annotating each noun phrase in a sentence with the corresponding boundin

g box in one of the frames of a video. Our dataset, ActivityNet-Entities, augmen ts the challenging ActivityNet Captions dataset with 158k bounding box annotatio ns, each grounding a noun phrase. This allows training video description models with this data, and importantly, evaluate how grounded or "true" such model are to the video they describe. To generate grounded captions, we propose a novel vi deo description model which is able to exploit these bounding box annotations. W e demonstrate the effectiveness of our model on our dataset, but also show how i t can be applied to image description on the Flickr30k Entities dataset. We achi eve state-of-the-art performance on video description, video paragraph descripti on, and image description and demonstrate our generated sentences are better gro unded in the video.

****

Streamlined Dense Video Captioning

Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, Bohyung Han; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p p. 6588-6597

Dense video captioning is an extremely challenging task since accurate and coher ent description of events in a video requires holistic understanding of video co ntents as well as contextual reasoning of individual events. Most existing appro aches handle this problem by first detecting event proposals from a video and th en captioning on a subset of the proposals. As a result, the generated sentences are prone to be redundant or inconsistent since they fail to consider temporal dependency between events. To tackle this challenge, we propose a novel dense vi deo captioning framework, which models temporal dependency across events in a vi deo explicitly and leverages visual and linguistic context from prior events for coherent storytelling. This objective is achieved by 1) integrating an event se quence generation network to select a sequence of event proposals adaptively, an d 2) feeding the sequence of event proposals to our sequential video captioning network, which is trained by reinforcement learning with two-level rewards---at both event and episode levels---for better context modeling. The proposed techni que achieves outstanding performances on ActivityNet Captions dataset in most me trics.

****

Adversarial Inference for Multi-Sentence Video Description

Jae Sung Park, Marcus Rohrbach, Trevor Darrell, Anna Rohrbach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 , pp. 6598-6608

While significant progress has been made in the image captioning task, video des cription is still in its infancy due to the complex nature of video data. Genera ting multi-sentence descriptions for long videos is even more challenging. Among the main issues are the fluency and coherence of the generated descriptions, an d their relevance to the video. Recently, reinforcement and adversarial learning based methods have been explored to improve the image captioning models; howeve r, both types of methods suffer from a number of issues, e.g. poor readability a nd high redundancy for RL and stability issues for GANs. In this work, we instea d propose to apply adversarial techniques during inference, designing a discrimi nator which encourages better multi-sentence video description. In addition, we find that a multi-discriminator "hybrid" design, where each discriminator target s one aspect of a description, leads to the best results. Specifically, we decou ple the discriminator to evaluate on three criteria: 1) visual relevance to the video, 2) language diversity and fluency, and 3) coherence across sentences. Our approach results in more accurate, diverse, and coherent multi-sentence video d escriptions, as shown by automatic as well as human evaluation on the popular Ac tivityNet Captions dataset.

****

Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, Wei-Y ing Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2019, pp. 6609-6618

We propose the Unified Visual-Semantic Embeddings (Unified VSE) for learning a joint space of visual representation and textual semantics. The model unifies the embeddings of concepts at different levels: objects, attributes, relations, and full scenes. We view the sentential semantics as a combination of different semantic components such as objects and relations; their embeddings are aligned with different image regions. A contrastive learning approach is proposed for the effective learning of this fine-grained alignment from only image-caption pairs. We also present a simple yet effective approach that enforces the coverage of caption embeddings on the semantic components that appear in the sentence. We demonstrate that the Unified VSE outperforms baselines on cross-modal retrieval tasks; the enforcement of the semantic coverage improves the model's robustness in defending text-domain adversarial attacks. Moreover, our model empowers the use of visual cues to accurately resolve word dependencies in novel sentences.
****

Learning to Compose Dynamic Tree Structures for Visual Contexts
Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6619-6628
We propose to compose dynamic tree structures that place the objects in an image into a visual context, helping visual reasoning tasks such as scene graph generation and visual Q&A. Our visual context tree model, dubbed VCTree, has two key advantages over existing structured object representations including chains and fully-connected graphs: 1) The efficient and expressive binary tree encodes the inherent parallel/hierarchical relationships among objects, e.g., "clothes" and "pants" are usually co-occur and belong to "person"; 2) the dynamic structure varies from image to image and task to task, allowing more content-/task-specific message passing among objects. To construct a VCTree, we design a score function that calculates the task-dependent validity between each object pair, and the tree is the binary version of the maximum spanning tree from the score matrix. Then, visual contexts are encoded by bidirectional TreeLSTM and decoded by task-specific models. We develop a hybrid learning procedure which integrates end-task supervised learning and the tree structure reinforcement learning, where the former's evaluation result serves as a self-critic for the latter's structure exploration. Experimental results on two benchmarks, which require reasoning over contexts: Visual Genome for scene graph generation and VQA2.0 for visual Q&A, show that VCTree outperforms state-of-the-art results while discovering interpretable visual context structures.
****

Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation
Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6629-6638
Vision-language navigation (VLN) is the task of navigating an embodied agent to carry out natural language instructions inside real 3D environments. In this paper, we study how to address three critical challenges for this task: the cross-modal grounding, the ill-posed feedback, and the generalization problems. First, we propose a novel Reinforced Cross-Modal Matching (RCM) approach that enforces cross-modal grounding both locally and globally via reinforcement learning (RL). Particularly, a matching critic is used to provide an intrinsic reward to encourage global matching between instructions and trajectories, and a reasoning navigator is employed to perform cross-modal grounding in the local visual scene. Evaluation on a VLN benchmark dataset shows that our RCM model significantly outperforms previous methods by 10% on SPL and achieves the new state-of-the-art performance. To improve the generalizability of the learned policy, we further introduce a Self-Supervised Imitation Learning (SIL) method to explore unseen environments by imitating its own past, good decisions. We demonstrate that SIL can approximate a better and more efficient policy, which tremendously minimizes the success rate performance gap between seen and unseen environments (from 30.7% to 11.7%).

****
Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering
Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6639-6648
Learning effective fusion of multi-modality features is at the heart of visual question answering. We propose a novel method of dynamically fuse multi-modal features with intra- and inter-modality information flow, which alternatively pass dynamic information between and across the visual and language modalities. It can robustly capture the high-level interactions between language and vision domains, thus significantly improves the performance of visual question answering. We also show that, the proposed dynamic intra modality attention flow conditioned on the other modality can dynamically modulate the intra-modality attention of the current modality, which is vital for multimodality feature fusion. Experimental evaluations on the VQA 2.0 dataset show that the proposed method achieves the state-of-the-art VQA performance. Extensive ablation studies are carried out for the comprehensive analysis of the proposed method.
****
Cycle-Consistency for Robust Visual Question Answering
Meet Shah, Xinlei Chen, Marcus Rohrbach, Devi Parikh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6649-6658
Despite significant progress in Visual Question Answer-ing over the years, robustness of today's VQA models leave much to be desired. We introduce a new evaluation protocol and associated dataset (VQA-Rephrasings) and show that state-of-the-art VQA models are notoriously brittle to linguistic variations in questions. VQA-Rephrasings contains 3 human-provided rephrasings for 40k questions-image pairs from the VQA v2.0 validation dataset. As a step towards improving robustness of VQA models, we propose a model-agnostic framework that exploits cycle consistency. Specifically, we train a model to not only answer a question, but also generate a question conditioned on the answer, such that the answer predicted for the generated question is the same as the ground truth answer to the original question. Without the use of additional supervision, we show that our approach is significantly more robust to linguistic variations than state-of-the-art VQA models, when evaluated on the VQA-Rephrasings dataset. In addition, our approach also outperforms state-of-the-art approaches on the standard VQA and Visual Question Generation tasks on the challenging VQA v2.0 dataset. Code and models will be made publicly available.
****
Embodied Question Answering in Photorealistic Environments With Point Cloud Perception
Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, Dhruv Batra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6659-6668
To help bridge the gap between internet vision-style problems and the goal of vision for embodied perception we instantiate a large-scale navigation task -- Embodied Question Answering [1] in photo-realistic environments (Matterport 3D). We thoroughly study navigation policies that utilize 3D point clouds, RGB images, or their combination. Our analysis of these models reveals several key findings. We find that two seemingly naive navigation baselines, forward-only and random, are strong navigators and challenging to outperform, due to the specific choice of the evaluation setting presented by [1]. We find a novel loss-weighting scheme we call Inflection Weighting to be important when training recurrent models for navigation with behavior cloning and are able to out perform the baselines with this technique. We find that point clouds provide a richer signal than RGB images for learning obstacle avoidance, motivating the use (and continued study) of 3D deep learning models for embodied navigation.
****

Reasoning Visual Dialogs With Structural and Partial Observations
Zilong Zheng, Wenguan Wang, Siyuan Qi, Song-Chun Zhu; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 666 9-6678
We propose a novel model to address the task of Visual Dialog which exhibits com plex dialog structures. To obtain a reasonable answer based on the current quest ion and the dialog history, the underlying semantic dependencies between dialog entities are essential. In this paper, we explicitly formalize this task as infe rence in a graphical model with partially observed nodes and unknown graph struc tures (relations in dialog). The given dialog entities are viewed as the observe d nodes. The answer to a given question is represented by a node with missing va lue. We first introduce an Expectation Maximization algorithm to infer both the underlying dialog structures and the missing node values (desired answers). Base d on this, we proceed to propose a differentiable graph neural network (GNN) sol ution that approximates this process. Experiment results on the VisDial and Vis Dial-Q datasets show that our model outperforms comparative methods. It is also observed that our method can infer the underlying dialog structure for better di alog reasoning.
****

Recursive Visual Attention in Visual Dialog
Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, Ji-Rong We n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2019, pp. 6679-6688
Visual dialog is a challenging vision-language task, which requires the agent to answer multi-round questions about an image. It typically needs to address two major problems: (1) How to answer visually-grounded questions, which is the core challenge in visual question answering (VQA); (2) How to infer the co-reference between questions and the dialog history. An example of visual co-reference is: pronouns (e.g., "they") in the question (e.g., "Are they on or off?") are linke d with nouns (e.g., "lamps") appearing in the dialog history (e.g., "How many la mps are there?") and the object grounded in the image. In this work, to resolve the visual co-reference for visual dialog, we propose a novel attention mechanis m called Recursive Visual Attention (RvA). Specifically, our dialog agent browse s the dialog history until the agent has sufficient confidence in the visual co- reference resolution, and refines the visual attention recursively. The quantita tive and qualitative experimental results on the large-scale VisDial v0.9 and v1 .0 datasets demonstrate that the proposed RvA not only outperforms the state-of- the-art methods, but also achieves reasonable recursion and interpretable attent ion maps without additional annotations. The code is available at https://github .com/yuleiniu/rva.
****

Two Body Problem: Collaborative Visual Task Completion
Unnat Jain, Luca Weihs, Eric Kolve, Mohammad Rastegari, Svetlana Lazebnik, Ali Farhadi, Alexander G. Schwing, Aniruddha Kembhavi; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 668 9-6699
Collaboration is a necessary skill to perform tasks that are beyond one agent's capabilities. Addressed extensively in both conventional and modern AI, multi-ag ent collaboration has often been studied in the context of simple grid worlds. W e argue that there are inherently visual aspects to collaboration which should b e studied in visually rich environments. A key element in collaboration is commu nication that can be either explicit, through messages, or implicit, through per ception of the other agents and the visual world. Learning to collaborate in a v isual environment entails learning (1) to perform the task, (2) when and what to communicate, and (3) how to act based on these communications and the perceptio n of the visual world. In this paper we study the problem of learning to collabo rate directly from pixels in AI2-THOR and demonstrate the benefits of explicit a nd implicit modes of communication to perform visual tasks. Refer to our project page for more details: https://prior.allenai.org/projects/two-body-problem
****

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question An swering
Drew A. Hudson, Christopher D. Manning; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6700-6709
We introduce GQA, a new dataset for real-world visual reasoning and compositiona l question answering, seeking to address key shortcomings of previous VQA datase ts. We have developed a strong and robust question engine that leverages Visual Genome scene graph structures to create 22M diverse reasoning questions, which a ll come with functional programs that represent their semantics. We use the prog rams to gain tight control over the answer distribution and present a new tunabl e smoothing technique to mitigate question biases. Accompanying the dataset is a suite of new metrics that evaluate essential qualities such as consistency, gro unding and plausibility. A careful analysis is performed for baselines as well a s state-of-the-art models, providing fine-grained results for different question types and topologies. Whereas a blind LSTM obtains a mere 42.1%, and strong VQA models achieve 54.1%, human performance tops at 89.3%, offering ample opportuni ty for new research to explore. We hope GQA will provide an enabling resource fo r the next generation of models with enhanced robustness, improved consistency, and deeper semantic understanding of vision and language.
****

Text2Scene: Generating Compositional Scenes From Textual Descriptions
Fuwen Tan, Song Feng, Vicente Ordonez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6710-6719
In this paper, we propose Text2Scene, a model that generates various forms of co mpositional scene representations from natural language descriptions. Unlike rec ent works, our method does NOT use Generative Adversarial Networks (GANs). Text2 Scene instead learns to sequentially generate objects and their attributes (loca tion, size, appearance, etc) at every time step by attending to different parts of the input text and the current status of the generated scene. We show that un der minor modifications, the proposed framework can handle the generation of dif ferent forms of scene representations, including cartoon-like scenes, object lay outs corresponding to real images, and synthetic images. Our method is not only competitive when compared with state-of-the-art GAN-based methods using automati c metrics and superior based on human judgments but also has the advantage of pr oducing interpretable results.
****

From Recognition to Cognition: Visual Commonsense Reasoning
Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 672 0-6731
Visual understanding goes well beyond object recognition. With one glance at an image, we can effortlessly imagine the world beyond the pixels: for instance, we can infer people's actions, goals, and mental states. While this task is easy f or humans, it is tremendously difficult for today's vision systems, requiring hi gher-order cognition and commonsense reasoning about the world. We formalize thi s task as Visual Commonsense Reasoning. Given a challenging question about an im age, a machine must answer correctly and then provide a rationale justifying its answer. Next, we introduce a new dataset, VCR, consisting of 290k multiple cho ice QA problems derived from 110k movie scenes. The key recipe for generating no n-trivial and high-quality problems at scale is Adversarial Matching, a new appr oach to transform rich annotations into multiple choice questions with minimal b ias. Experimental results show that while humans find VCR easy (over 90% accurac y), state-of-the-art vision models struggle (45%). To move towards cognition-l evel understanding, we present a new reasoning engine, Recognition to Cognition Networks (R2C), that models the necessary layered inferences for grounding, cont extualization, and reasoning. R2C helps narrow the gap between humans and machin es (65%); still, the challenge is far from solved, and we provide analysis that suggests avenues for future work.
****

The Regretful Agent: Heuristic-Aided Navigation Through Progress Estimation

Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6732-6740

As deep learning continues to make progress for challenging perception tasks, there is increased interest in combining vision, language, and decision-making. Specifically, the Vision and Language Navigation (VLN) task involves navigating to a goal purely from language instructions and visual information without explicit knowledge of the goal. Recent successful approaches have made in-roads in achieving good success rates for this task but rely on beam search, which thoroughly explores a large number of trajectories and is unrealistic for applications such as robotics. In this paper, inspired by the intuition of viewing the problem as search on a navigation graph, we propose to use a progress monitor developed in prior work as a learnable heuristic for search. We then propose two modules incorporated into an end-to-end architecture: 1) A learned mechanism to perform backtracking, which decides whether to continue moving forward or roll back to a previous state (Regret Module) and 2) A mechanism to help the agent decide which direction to go next by showing directions that are visited and their associated progress estimate (Progress Marker). Combined, the proposed approach significantly outperforms current state-of-the-art methods using greedy action selection, with 5% absolute improvement on the test server in success rates, and more importantly 8% on success rates normalized by the path length.

****

Tactical Rewind: Self-Correction via Backtracking in Vision-And-Language Navigation

Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, Siddhartha Srinivasa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6741-6749

We present the Frontier Aware Search with backTracking (FAST) Navigator, a general framework for action decoding, that achieves state-of-the-art results on the 2018 Room-to-Room (R2R) Vision-and-Language navigation challenge. Given a natural language instruction and photo-realistic image views of a previously unseen environment, the agent was tasked with navigating from source to target location as quickly as possible. While all current approaches make local action decisions or score entire trajectories using beam search, ours balances local and global signals when exploring an unobserved environment. Importantly, this lets us act greedily but use global signals to backtrack when necessary. Applying FAST framework to existing state-of-the-art models achieved a 17% relative gain, an absolute 6% gain on Success rate weighted by Path Length.

****

Learning to Learn How to Learn: Self-Adaptive Visual Navigation Using Meta-Learning

Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, Roozbeh Mottaghi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6750-6759

Learning is an inherently continuous phenomenon. When humans learn a new task there is no explicit distinction between training and inference. As we learn a task, we keep learning about it while performing the task. What we learn and how we learn it varies during different stages of learning. Learning how to learn and adapt is a key property that enables us to generalize effortlessly to new settings. This is in contrast with conventional settings in machine learning where a trained model is frozen during inference. In this paper we study the problem of learning to learn at both training and test time in the context of visual navigation. A fundamental challenge in navigation is generalization to unseen scenes. In this paper we propose a self-adaptive visual navigation method (SAVN) which learns to adapt to new environments without any explicit supervision. Our solution is a meta-reinforcement learning approach where an agent learns a self-supervised interaction loss that encourages effective navigation. Our experiments, performed in the AI2-THOR framework, show major improvements in both success rate and SPL for visual navigation in novel scenes. Our code and data are available at: https://github.com/allenai/savn.

****
High Flux Passive Imaging With Single-Photon Sensors
Atul Ingle, Andreas Velten, Mohit Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6760-6769
Single-photon avalanche diodes (SPADs) are an emerging technology with a unique capability of capturing individual photons with high timing precision. SPADs are being used in several active imaging systems (e.g., fluorescence lifetime microscopy and LiDAR), albeit mostly limited to low photon flux settings. We propose passive free-running SPAD (PF-SPAD) imaging, an imaging modality that uses SPADs for capturing 2D intensity images with unprecedented dynamic range under ambient lighting, without any active light source. Our key observation is thatthe precise inter-photon timing measured by a SPAD can be used for estimating scene brightness under ambient lighting conditions, even for very bright scenes. We develop a theoretical model for PF-SPAD imaging, and derive a scene brightness estimator based on the average time of darkness between successive photons detected by a PF-SPAD pixel. Our key insight is that due to the stochastic nature of photon arrivals, this estimator does not suffer from a hard saturation limit. Coupled with high sensitivity at low flux, this enables a PF-SPAD pixel to measure a wide range of scene brightnesses, from very low to very high, thereby achieving extreme dynamic range. We demonstrate an improvement of over 2 orders of magnitude over conventional sensors by imaging scenes spanning a dynamic range of 10^6:1.
****
Photon-Flooded Single-Photon 3D Cameras
Anant Gupta, Atul Ingle, Andreas Velten, Mohit Gupta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6770-6779
Single-photon avalanche diodes (SPADs) are starting to play a pivotal role in the development of photon-efficient, long-range LiDAR systems. However, due to non-linearities in their image formation model, a high photon flux (e.g., due to strong sunlight) leads to distortion of the incident temporal waveform, and potentially, large depth errors. Operating SPADs in low flux regimes can mitigate these distortions, but, often requires attenuating the signal and thus, results in low signal-to-noise ratio. In this paper, we address the following basic question: what is the optimal photon flux that a SPAD-based LiDAR should be operated in? We derive a closed form expression for the optimal flux, which is quasi-depth-invariant, and depends on the ambient light strength. The optimal flux is lower than what a SPAD typically measures in real world scenarios, but surprisingly, considerably higher than what is conventionally suggested for avoiding distortions. We propose a simple, adaptive approach for achieving the optimal flux by attenuating incident flux based on an estimate of ambient light strength. Using extensive simulations and a hardware prototype, we show that the optimal flux criterion holds for several depth estimators, under a wide range of illumination conditions.
****
Acoustic Non-Line-Of-Sight Imaging
David B. Lindell, Gordon Wetzstein, Vladlen Koltun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6780-6789
Non-line-of-sight (NLOS) imaging enables unprecedented capabilities in a wide range of applications, including robotic and machine vision, remote sensing, autonomous vehicle navigation, and medical imaging. Recent approaches to solving this challenging problem employ optical time-of-flight imaging systems with highly sensitive time-resolved photodetectors and ultra-fast pulsed lasers. However, despite recent successes in NLOS imaging using these systems, widespread implementation and adoption of the technology remains a challenge because of the requirement for specialized, expensive hardware. We introduce acoustic NLOS imaging, which is orders of magnitude less expensive than most optical systems and captures hidden 3D geometry at longer ranges with shorter acquisition times compared to state-of-the-art optical methods. Inspired by hardware setups used in radar and algorithmic approaches to model and invert wave-based image formation models devel

oped in the seismic imaging community, we demonstrate a new approach to seeing a round corners.
****

Steady-State Non-Line-Of-Sight Imaging
Wenzheng Chen, Simon Daneau, Fahim Mannan, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6790-6799
Conventional intensity cameras recover objects in the direct line-of-sight of the camera, whereas occluded scene parts are considered lost in this process. Non-line-of-sight imaging (NLOS) aims at recovering these occluded objects by analyzing their indirect reflections on visible scene surfaces. Existing NLOS methods temporally probe the indirect light transport to unmix light paths based on their travel time, which mandates specialized instrumentation that suffers from low photon efficiency, high cost, and mechanical scanning. We depart from temporal probing and demonstrate steady-state NLOS imaging using conventional intensity sensors and continuous illumination. Instead of assuming perfectly isotropic scattering, the proposed method exploits directionality in the hidden surface reflectance, resulting in (small) spatial variation of their indirect reflections for varying illumination. To tackle the shape-dependence of these variations, we propose a trainable architecture which learns to map diffuse indirect reflections to scene reflectance using only synthetic training data. Relying on consumer color image sensors, with high fill factor, high quantum efficiency and low read-out noise, we demonstrate high-fidelity color NLOS imaging for scene configurations tackled before with picosecond time resolution.
****

A Theory of Fermat Paths for Non-Line-Of-Sight Shape Reconstruction
Shumian Xin, Sotiris Nousias, Kiriakos N. Kutulakos, Aswin C. Sankaranarayanan, Srinivasa G. Narasimhan, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6800-6809
We present a novel theory of Fermat paths of light between a known visible scene and an unknown object not in the line of sight of a transient camera. These light paths either obey specular reflection or are reflected by the object's boundary, and hence encode the shape of the hidden object. We prove that Fermat paths correspond to discontinuities in the transient measurements. We then derive a novel constraint that relates the spatial derivatives of the path lengths at these discontinuities to the surface normal. Based on this theory, we present an algorithm, called Fermat Flow, to estimate the shape of the non-line-of-sight object. Our method allows, for the first time, accurate shape recovery of complex objects, ranging from diffuse to specular, that are hidden around the corner as well as hidden behind a diffuser. Finally, our approach is agnostic to the particular technology used for transient imaging. As such, we demonstrate mm-scale shape recovery from pico-second scale transients using a SPAD and ultrafast laser, as well as micron-scale reconstruction from femto-second scale transients using interferometry. We believe our work is a significant advance over the state-of-the-art in non-line-of-sight imaging.
****

End-To-End Projector Photometric Compensation
Bingyao Huang, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6810-6819
Projector photometric compensation aims to modify a projector input image such that it can compensate for disturbance from the appearance of projection surface. In this paper, for the first time, we formulate the compensation problem as an end-to-end learning problem and propose a convolutional neural network, named CompenNet, to implicitly learn the complex compensation function. CompenNet consists of a UNet-like backbone network and an autoencoder subnet. Such architecture encourages rich multi-level interactions between the camera-captured projection surface image and the input image, and thus captures both photometric and environment information of the projection surface. In addition, the visual details and interaction information are carried to deeper layers along the multi-level skip convolution layers. The architecture is of particular importance for the proje

ctor compensation task, for which only a small training dataset is allowed in pr
actice. Another contribution we make is a novel evaluation benchmark, which is i
ndependent of system setup and thus quantitatively verifiable. Such benchmark is
 not previously available, to our best knowledge, due to the fact that conventio
nal evaluation requests the hardware system to actually project the final result
s. Our key idea, motivated from our end-to-end problem formulation, is to use a
reasonable surrogate to avoid such projection process so as to be setup-independ
ent. Our method is evaluated carefully on the benchmark, and the results show th
at our end-to-end learning solution outperforms state-of-the-arts both qualitati
vely and quantitatively by a significant margin.
****

Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera
Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, Yuch
ao Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2019, pp. 6820-6829
Event-based cameras can measure intensity changes (called 'events') with microse
cond accuracy under high-speed motion and challenging lighting conditions. With
the active pixel sensor (APS), the event camera allows simultaneous output of th
e intensity frames. However, the output images are captured at a relatively low
frame-rate and often suffer from motion blur. A blurry image can be regarded as
the integral of a sequence of latent images, while the events indicate the chang
es between the latent images. Therefore, we are able to model the blur-generatio
n process by associating event data to a latent image. In this paper, we propose
 a simple and effective approach, the Event-based Double Integral (EDI) model, t
o reconstruct a high frame-rate, sharp video from a single blurry frame and its
event data. The video generation is based on solving a simple non-convex optimiz
ation problem in a single scalar variable. Experimental results on both syntheti
c and real images demonstrate the superiority of our EDI model and optimization
method in comparison to the state-of-the-art.
****

Bringing Alive Blurred Moments
Kuldeep Purohit, Anshul Shah, A. N. Rajagopalan; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6830-6839
We present a solution for the goal of extracting a video from a single motion bl
urred image to sequentially reconstruct the clear views of a scene as beheld by
the camera during the time of exposure. We first learn motion representation fro
m sharp videos in an unsupervised manner through training of a convolutional rec
urrent video autoencoder network that performs a surrogate task of video reconst
ruction. Once trained, it is employed for guided training of a motion encoder fo
r blurred images. This network extracts embedded motion information from the blu
rred image to generate a sharp video in conjunction with the trained recurrent v
ideo decoder. As an intermediate step, we also design an efficient architecture
that enables real-time single image deblurring and outperforms competing methods
 across all factors: accuracy, speed, and compactness. Experiments on real scene
s and standard datasets demonstrate the superiority of our framework over the st
ate-of-the-art and its ability to generate a plausible sequence of temporally co
nsistent sharp frames.
****

Learning to Synthesize Motion Blur
Tim Brooks, Jonathan T. Barron; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2019, pp. 6840-6848
We present a technique for synthesizing a motion blurred image from a pair of un
blurred images captured in succession. To build this system we motivate and desi
gn a differentiable "line prediction" layer to be used as part of a neural netwo
rk architecture, with which we can learn a system to regress from image pairs to
 motion blurred images that span the capture time of the input image pair. Train
ing this model requires an abundance of data, and so we design and execute a str
ategy for using frame interpolation techniques to generate a large-scale synthet
ic dataset of motion blurred images and their respective inputs. We additionally
 capture a high quality test set of real motion blurred images, synthesized from

slow motion videos, with which we evaluate our model against several baseline techniques that can be used to synthesize motion blur. Our model produces higher accuracy output than our baselines, and is several orders of magnitude faster than baselines with competitive accuracy.
****
Underexposed Photo Enhancement Using Deep Illumination Estimation
Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6849-6857
This paper presents a new neural network for enhancing underexposed photos. Instead of directly learning an image-to-image mapping as previous work, we introduce intermediate illumination in our network to associate the input with expected enhancement result, which augments the network's capability to learn complex photographic adjustment from expert-retouched input/output image pairs. Based on this model, we formulate a loss function that adopts constraints and priors on the illumination, prepare a new dataset of 3,000 underexposed image pairs, and train the network to effectively learn a rich variety of adjustment for diverse lighting conditions. By these means, our network is able to recover clear details, distinct contrast, and natural color in the enhancement results. We perform extensive experiments on the benchmark MIT-Adobe FiveK dataset and our new dataset, and show that our network is effective to deal with previously challenging images.
****
Blind Visual Motif Removal From a Single Image
Amir Hertz, Sharon Fogel, Rana Hanocka, Raja Giryes, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6858-6867
Many images shared over the web include overlaid objects, or visual motifs, such as text, symbols or drawings, which add a description or decoration to the image. For example, decorative text that specifies where the image was taken, repeatedly appears across a variety of different images. Often, the reoccurring visual motif, is semantically similar, yet, differs in location, style and content (e.g. text placement, font and letters). This work proposes a deep learning based technique for blind removal of such objects. In the blind setting, the location and exact geometry of the motif are unknown. Our approach simultaneously estimates which pixels contain the visual motif, and synthesizes the underlying latent image. It is applied to a single input image, without any user assistance in specifying the location of the motif, achieving state-of-the-art results for blind removal of both opaque and semi-transparent visual motifs.
****
Non-Local Meets Global: An Integrated Paradigm for Hyperspectral Denoising
Wei He, Quanming Yao, Chao Li, Naoto Yokoya, Qibin Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6868-6877
Non-local low-rank tensor approximation has been developed as a state-of-the-art method for hyperspectral image (HSI) denoising. Unfortunately, while their denoising performance benefits little from more spectral bands, the running time of these methods significantly increases. In this paper, we claim that the HSI lies in a global spectral low-rank subspace, and the spectral subspaces of each full band patch groups should lie in this global low-rank subspace. This motivates us to propose a unified spatial-spectral paradigm for HSI denoising. As the new model is hard to optimize, An efficient algorithm motivated by alternating minimization is developed. This is done by first learning a low-dimensional orthogonal basis and the related reduced image from the noisy HSI. Then, the non-local low-rank denoising and iterative regularization are developed to refine the reduced image and orthogonal basis, respectively. Finally, the experiments on synthetic and both real datasets demonstrate the superiority against the
****
Neural Rerendering in the Wild
Moustafa Meshry, Dan B. Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey,

Noah Snavely, Ricardo Martin-Brualla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6878-6887
We explore total scene capture --- recording, modeling, and rerendering a scene under varying appearance such as season and time of day. Starting from Internet photos of a tourist landmark, we apply traditional 3D reconstruction to register the photos and approximate the scene as a point cloud. For each photo, we render the scene points into a deep framebuffer, and train a deep neural network to learn the mapping of these initial renderings to the actual photos. This rendering network also takes as input a latent appearance vector and a semantic mask indicating the location of transient objects like pedestrians. The model is evaluated on several datasets of publicly available images spanning a broad range of illumination conditions. We create short videos that demonstrate realistic manipulation of the image viewpoint, appearance, and semantic labels. We also compare results to prior work on scene reconstruction from Internet photos.
****

GeoNet: Deep Geodesic Networks for Point Cloud Analysis
Tong He, Haibin Huang, Li Yi, Yuqian Zhou, Chihao Wu, Jue Wang, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6888-6897
Surface-based geodesic topology provides strong cues for object semantic analysis and geometric modeling. However, such connectivity information is lost in point clouds. Thus we introduce GeoNet, the first deep learning architecture trained to model the intrinsic structure of surfaces represented as point clouds. To demonstrate the applicability of learned geodesic-aware representations, we propose fusion schemes which use GeoNet in conjunction with other baseline or backbone networks, such as PU-Net and PointNet++, for down-stream point cloud analysis. Our method improves the state-of-the-art on multiple representative tasks that can benefit from understandings of the underlying surface topology, including point upsampling, normal estimation, mesh reconstruction and non-rigid shape classification.
****

MeshAdv: Adversarial Meshes for Visual Recognition
Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, Mingyan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6898-6907
Highly expressive models such as deep neural networks (DNNs) have been widely applied to various applications. However, recent studies show that DNNs are vulnerable to adversarial examples, which are carefully crafted inputs aiming to mislead the predictions. Currently, the majority of these studies have focused on perturbation added to image pixels, while such manipulation is not physically realistic. Some works have tried to overcome this limitation by attaching printable 2D patches or painting patterns onto surfaces, but can be potentially defended because 3D shape features are intact. In this paper, we propose meshAdv to generate "adversarial 3D meshes" from objects that have rich shape features but minimal textural variation. To manipulate the shape or texture of the objects, we make use of a differentiable renderer to compute accurate shading on the shape and propagate the gradient. Extensive experiments show that the generated 3D meshes are effective in attacking both classifiers and object detectors. We evaluate the attack under different viewpoints. In addition, we design a pipeline to perform black-box attack on a photorealistic renderer with unknown rendering parameters.
****

Fast Spatially-Varying Indoor Lighting Estimation
Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Jean-Francois Lalonde; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6908-6917
We propose a real-time method to estimate spatially-varying indoor lighting from a single RGB image. Given an image and a 2D location in that image, our CNN estimates a 5th order spherical harmonic representation of the lighting at the given location in less than 20ms on a laptop mobile graphics card. While existing approaches estimate a single, global lighting representation or require depth as i

nput, our method reasons about local lighting without requiring any geometry inf
ormation. We demonstrate, through quantitative experiments including a user stud
y, that our results achieve lower lighting estimation errors and are preferred b
y users over the state-of-the-art. Our approach can be used directly for augment
ed reality applications, where a virtual object is relit realistically at any po
sition in the scene in real-time.
****

Neural Illumination: Lighting Prediction for Indoor Environments
Shuran Song, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2019, pp. 6918-6926
This paper addresses the task of estimating the light arriving from all directio
ns to a 3D point observed at a selected pixel in an RGB image. This task is ch
allenging because it requires predicting a mapping from a partial RGB observatio
n by a camera to a complete illumination map for a different 3D point, which dep
ends on the 3D location of the selected pixel, the distribution of unobserved li
ght sources, the occlusions by scene geometry, etc. Previous methods attempt to
learn this complex mapping directly using a single black-box neural network whi
ch often fails to estimate high-frequency lighting details for scenes with compl
icated 3D geometry. Instead, we propose "Neural Illumination," a new approach t
hat decomposes illumination prediction into several simpler differentiable sub-t
asks: 1) geometry estimation, 2) scene completion, and 3) LDR-to-HDR estimation.
The advantage of this approach is that the sub-tasks are relatively easy to le
arn and can be trained with direct supervision, while the whole pipeline is full
y differentiable and can be fine-tuned with end-to-end supervision. Experiment
s show that our approach performs significantly better quantitatively and qualit
atively than prior work.
****

Deep Sky Modeling for Single Image Outdoor Lighting Estimation
Yannick Hold-Geoffroy, Akshaya Athawale, Jean-Francois Lalonde; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019
, pp. 6927-6935
We propose a data-driven learned sky model, which we use for outdoor lighting es
timation from a single image. As no large-scale dataset of images and their corr
esponding ground truth illumination is readily available, we use complementary d
atasets to train our approach, combining the vast diversity of illumination cond
itions of SUN360 with the radiometrically calibrated and physically accurate Lav
al HDR sky database. Our key contribution is to provide a holistic view of both
lighting modeling and estimation, solving both problems end-to-end. From a test
image, our method can directly estimate an HDR environment map of the lighting w
ithout relying on analytical lighting models. We demonstrate the versatility and
expressivity of our learned sky model and show that it can be used to recover p
lausible illumination, leading to visually pleasant virtual object insertions. T
o further evaluate our method, we capture a dataset of HDR 360deg panoramas and
show through extensive validation that we significantly outperform previous stat
e-of-the-art.
****

Bidirectional Learning for Domain Adaptation of Semantic Segmentation
Yunsheng Li, Lu Yuan, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6936-6945
Domain adaptation for semantic image segmentation is very necessary since manual
ly labeling large datasets with pixel-level labels is expensive and time consumi
ng. Existing domain adaptation techniques either work on limited datasets, or yi
eld not so good performance compared with supervised learning. In this paper, w
e propose a novel bidirectional learning framework for domain adaptation of segm
entation. Using the bidirectional learning, the image translation model and the
segmentation adaptation model can be learned alternatively and promote to each o
ther.Furthermore, we propose a self-supervised learning algorithm to learn a bet
ter segmentation adaptation model and in return improve the image translation mo
del. Experiments show that our method superior to the state-of-the-art methods i
n domain adaptation of segmentation with a big margin. The source code is availa

ble at https://github.com/liyunsheng13/BDL
****
Enhanced Bayesian Compression via Deep Reinforcement Learning
Xin Yuan, Liangliang Ren, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6946-6955
In this paper, we propose an Enhanced Bayesian Compression method to flexibly compress the deep networks via reinforcement learning. Unlike the existing Bayesian compression method which cannot explicitly enforce quantization weights during training, our method learns flexible codebooks in each layer for an optimal network quantization. To dynamically adjust the state of codebooks, we employ an Actor-Critic network to collaborate with the original deep network. Different from most existing network quantization methods, our EBC does not require re-training procedures after the quantization. Experimental results show that our method obtains low-bit precision with acceptable accuracy drop on MNIST, CIFAR and ImageNet.
****
Strong-Weak Distribution Alignment for Adaptive Object Detection
Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, Kate Saenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6956-6965
We propose an approach for unsupervised adaptation of object detectors from label-rich to label-poor domains which can significantly reduce annotation costs associated with detection. Recently, approaches that align distributions of source and target images using an adversarial loss have been proven effective for adapting object classifiers. However, for object detection, fully matching the entire distributions of source and target images to each other at the global image level may fail, as domains could have distinct scene layouts and different combinations of objects. On the other hand, strong matching of local features such as texture and color makes sense, as it does not change category level semantics. This motivates us to propose a novel method for detector adaptation based on strong local alignment and weak global alignment. Our key contribution is the weak alignment model, which focuses the adversarial alignment loss on images that are globally similar and puts less emphasis on aligning images that are globally dissimilar. Additionally, we design the strong domain alignment model to only look at local receptive fields of the feature map. We empirically verify the effectiveness of our method on four datasets comprising both large and small domain shifts. Our code is available at https://github.com/VisionLearningGroup/DA_Detection.
****
MFAS: Multimodal Fusion Architecture Search
Juan-Manuel Perez-Rua, Valentin Vielzeuf, Stephane Pateux, Moez Baccouche, Frederic Jurie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6966-6975
We tackle the problem of finding good architectures for multimodal classification problems. We propose a novel and generic search space that spans a large number of possible fusion architectures. In order to find an optimal architecture for a given dataset in the proposed search space, we leverage an efficient sequential model-based exploration approach that is tailored for the problem. We demonstrate the value of posing multimodal fusion as a neural architecture search problem by extensive experimentation on a toy dataset and two other real multimodal datasets. We discover fusion architectures that exhibit state-of-the-art performance for problems with different domain and dataset size, including the \ntu dataset, the largest multimodal action recognition dataset available.
****
Disentangling Adversarial Robustness and Generalization
David Stutz, Matthias Hein, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6976-6987
Obtaining deep networks that are robust against adversarial examples and generalize well is an open problem. A recent hypothesis even states that both robust and accurate models are impossible, i.e., adversarial robustness and generalization are conflicting goals. In an effort to clarify the relationship between robust

ness and generalization, we assume an underlying, low-dimensional data manifold and show that: 1. regular adversarial examples leave the manifold; 2. adversarial examples constrained to the manifold, i.e., on-manifold adversarial examples, exist; 3. on-manifold adversarial examples are generalization errors, and on-manifold adversarial training boosts generalization; 4. regular robustness and generalization are not necessarily contradicting goals. These assumptions imply that both robust and accurate models are possible. However, different models (architectures, training strategies etc.) can exhibit different robustness and generalization characteristics. To confirm our claims, we present extensive experiments on synthetic data (with known manifold) as well as on EMNIST, Fashion-MNIST and CelebA.

****

ShieldNets: Defending Against Adversarial Attacks Using Probabilistic Adversarial Robustness

Rajkumar Theagarajan, Ming Chen, Bir Bhanu, Jing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6988-6996

Defending adversarial attack is a critical step towards reliable deployment of deep learning empowered solutions for industrial applications. Probabilistic adversarial robustness (PAR), as a theoretical framework, is introduced to neutralize adversarial attacks by concentrating sample probability to adversarial-free zones. Distinct to most of the existing defense mechanisms that require modifying the architecture/training of the target classifier which is not feasible in the real-world scenario, e.g., when a model has already been deployed, PAR is designed in the first place to provide proactive protection to an existing fixed model. ShieldNet is implemented as a demonstration of PAR in this work by using PixelCNN. Experimental results show that this approach is generalizable, robust against adversarial transferability and resistant to a wide variety of attacks on the Fashion-MNIST and CIFAR10 datasets, respectively.

****

Deeply-Supervised Knowledge Synergy

Dawei Sun, Anbang Yao, Aojun Zhou, Hao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6997-7006

Convolutional Neural Networks (CNNs) have become deeper and more complicated compared with the pioneering AlexNet. However, current prevailing training scheme follows the previous way of adding supervision to the last layer of the network only and propagating error information up layer-by-layer. In this paper, we propose Deeply-supervised Knowledge Synergy (DKS), a new method aiming to train CNNs with improved generalization ability for image classification tasks without introducing extra computational cost during inference. Inspired by the deeply-supervised learning scheme, we first append auxiliary supervision branches on top of certain intermediate network layers. While properly using auxiliary supervision can improve model accuracy to some degree, we go one step further to explore the possibility of utilizing the probabilistic knowledge dynamically learnt by the classifiers connected to the backbone network as a new regularization to improve the training. A novel synergy loss, which considers pairwise knowledge matching among all supervision branches, is presented. Intriguingly, it enables dense pairwise knowledge matching operations in both top-down and bottom-up directions at each training iteration, resembling a dynamic synergy process for the same task. We evaluate DKS on image classification datasets using state-of-the-art CNN architectures, and show that the models trained with it are consistently better than the corresponding counterparts. For instance, on the ImageNet classification benchmark, our ResNet-152 model outperforms the baseline model with a 1.47% margin in Top-1 accuracy. Code is available at https://github.com/sundw2014/DKS.

****

Dual Residual Networks Leveraging the Potential of Paired Operations for Image Restoration

Xing Liu, Masanori Suganuma, Zhun Sun, Takayuki Okatani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7007-7016

In this paper, we study design of deep neural networks for tasks of image restoration. We propose a novel style of residual connections dubbed "dual residual connection", which exploits the potential of paired operations, e.g., up- and down-sampling or convolution with large- and small-size kernels. We design a modular block implementing this connection style; it is equipped with two containers to which arbitrary paired operations are inserted. Adopting the "unraveled" view of the residual networks proposed by Veit et al., we point out that a stack of the proposed modular blocks allows the first operation in a block interact with the second operation in any subsequent blocks. Specifying the two operations in each of the stacked blocks, we build a complete network for each individual task of image restoration. We experimentally evaluate the proposed approach on five image restoration tasks using nine datasets. The results show that the proposed networks with properly chosen paired operations outperform previous methods on almost all of the tasks and datasets.

****

Probabilistic End-To-End Noise Correction for Learning With Noisy Labels

Kun Yi, Jianxin Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7017-7025

Deep learning has achieved excellent performance in various computer vision tasks, but requires a lot of training examples with clean labels. It is easy to collect a dataset with noisy labels, but such noise makes networks overfit seriously and accuracies drop dramatically. To address this problem, we propose an end-to-end framework called PENCIL, which can update both network parameters and label estimations as label distributions. PENCIL is independent of the backbone network structure and does not need an auxiliary clean dataset or prior information about noise, thus it is more general and robust than existing methods and is easy to apply. PENCIL outperformed previous state-of-the-art methods by large margins on both synthetic and real-world datasets with different noise types and noise rates. Experiments show that PENCIL is robust on clean datasets, too.

****

Attention-Guided Unified Network for Panoptic Segmentation

Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, Xinga ng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7026-7035

This paper studies panoptic segmentation, a recently proposed task which segments foreground (FG) objects at the instance level as well as background (BG) contents at the semantic level. Existing methods mostly dealt with these two problems separately, but in this paper, we reveal the underlying relationship between them, in particular, FG objects provide complementary cues to assist BG understanding. Our approach, named the Attention-guided Unified Network (AUNet), is a unified framework with two branches for FG and BG segmentation simultaneously. Two sources of attentions are added to the BG branch, namely, RPN and FG segmentation mask to provide object-level and pixel-level attentions, respectively. Our approach is generalized to different backbones with consistent accuracy gain in both FG and BG segmentation, and also sets new state-of-the-arts both in the MS-COCO (46.5% PQ) and Cityscapes (59.0% PQ) benchmarks.

****

NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection

Golnaz Ghiasi, Tsung-Yi Lin, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7036-7045

Current state-of-the-art convolutional architectures for object detection are manually designed. Here we aim to learn a better architecture of feature pyramid network for object detection. We adopt Neural Architecture Search and discover a new feature pyramid architecture in a novel scalable search space covering all cross-scale connections. The discovered architecture, named NAS-FPN, consists of a combination of top-down and bottom-up connections to fuse features across scales. NAS-FPN, combined with various backbone models in the RetinaNet framework, achieves better accuracy and latency tradeoff compared to state-of-the-art object detection models. NAS-FPN improves mobile detection accuracy by 2 AP compared to state-of-the-art SSDLite with MobileNetV2 model in [32] and achieves 48.3 AP w

hich surpasses Mask R-CNN [10] detection accuracy with less computation time.
****

OICSR: Out-In-Channel Sparsity Regularization for Compact Deep Neural Networks
Jiashi Li, Qi Qi, Jingyu Wang, Ce Ge, Yujian Li, Zhangzhang Yue, Haifeng S un; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2019, pp. 7046-7055
Channel pruning can significantly accelerate and compress deep neural networks. Many channel pruning works utilize structured sparsity regularization to zero ou t all the weights in some channels and automatically obtain structure-sparse net work in training stage. However, these methods apply structured sparsity regular ization on each layer separately where the correlations between consecutive laye rs are omitted. In this paper, we first combine one out-channel in current layer and the corresponding in-channel in next layer as a regularization group, namel y out-in-channel. Our proposed Out-In-Channel Sparsity Regularization (OICSR) co nsiders correlations between successive layers to further retain predictive powe r of the compact network. Training with OICSR thoroughly transfers discriminativ e features into a fraction of out-in-channels. Correspondingly, OICSR measures c hannel importance based on statistics computed from two consecutive layers, not individual layer. Finally, a global greedy pruning algorithm is designed to remo ve redundant out-in-channels in an iterative way. Our method is comprehensively evaluated with various CNN architectures including CifarNet, AlexNet, ResNet, De nseNet and PreActSeNet on CIFAR-10, CIFAR-100 and ImageNet-1K datasets. Notably, on ImageNet-1K, we reduce 37.2% FLOPs on ResNet-50 while outperforming the orig inal model by 0.22% top-1 accuracy.
****

Semantically Aligned Bias Reducing Zero Shot Learning
Akanksha Paul, Narayanan C. Krishnan, Prateek Munjal; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7056 -7065
Zero shot learning (ZSL) aims to recognize unseen classes by exploiting semantic relationships between seen and unseen classes. Two major problems faced by ZSL algorithms are the hubness problem and the bias towards the seen classes. Existi ng ZSL methods focus on only one of these problems in the conventional and gener alized ZSL setting. In this work, we propose a novel approach, Semantically Alig ned Bias Reducing (SABR) ZSL, which focuses on solving both the problems. It ove rcomes the hubness problem by learning a latent space that preserves the semanti c relationship between the labels while encoding the discriminating information about the classes. Further, we also propose ways to reduce bias of the seen cla sses through a simple cross-validation process in the inductive setting and a no vel weak transfer constraint in the transductive setting. Extensive experiments on three benchmark datasets suggest that the proposed model significantly outpe rforms existing state-of-the-art algorithms by 1.5-9% in the conventional ZSL s etting and by 2-14% in the generalized ZSL for both the inductive and transduct ive settings.
****

Feature Space Perturbations Yield More Transferable Adversarial Examples
Nathan Inkawhich, Wei Wen, Hai (Helen) Li, Yiran Chen; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 70 66-7074
Many recent works have shown that deep learning models are vulnerable to quasi-i mperceptible input perturbations, yet practitioners cannot fully explain this be havior. This work describes a transfer-based blackbox targeted adversarial attac k of deep feature space representations that also provides insights into cross-m odel class representations of deep CNNs. The attack is explicitly designed for t ransferability and drives feature space representation of a source image at laye r L towards the representation of a target image at L. The attack yields highly transferable targeted examples, which outperform competition winning methods by over 30% in targeted attack metrics. We also show the choice of L to generate ex amples from is important, transferability characteristics are blackbox model agn ostic, and indicate that well trained deep models have similar highly-abstract r

epresentations.
****

IGE-Net: Inverse Graphics Energy Networks for Human Pose Estimation and Single-View Reconstruction
Dominic Jack, Frederic Maire, Sareh Shirazi, Anders Eriksson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7075-7084
Inferring 3D scene information from 2D observations is an open problem in computer vision. We propose using a deep-learning based energy minimization framework to learn a consistency measure between 2D observations and a proposed world model, and demonstrate that this framework can be trained end-to-end to produce consistent and realistic inferences. We evaluate the framework on human pose estimation and voxel-based object reconstruction benchmarks and show competitive results can be achieved with relatively shallow networks with drastically fewer learned parameters and floating point operations than conventional deep-learning approaches.
****

Accelerating Convolutional Neural Networks via Activation Map Compression
Georgios Georgiadis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7085-7095
The deep learning revolution brought us an extensive array of neural network architectures that achieve state-of-the-art performance in a wide variety of Computer Vision tasks including among others, classification, detection and segmentation. In parallel, we have also been observing an unprecedented demand in computational and memory requirements, rendering the efficient use of neural networks in low-powered devices virtually unattainable. Towards this end, we propose a three-stage compression and acceleration pipeline that sparsifies, quantizes and entropy encodes activation maps of Convolutional Neural Networks. Sparsification increases the representational power of activation maps leading to both acceleration of inference and higher model accuracy. Inception-V3 and MobileNet-V1 can be accelerated by as much as 1.6x with an increase in accuracy of 0.38% and 0.54% on the ImageNet and CIFAR-10 datasets respectively. Quantizing and entropy coding the sparser activation maps lead to higher compression over the baseline, reducing the memory cost of the network execution. Inception-V3 and MobileNet-V1 activation maps, quantized to 16 bits, are compressed by as much as 6x with an increase in accuracy of 0.36% and 0.55% respectively.
****

Knowledge Distillation via Instance Relationship Graph
Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, Yunqiang Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7096-7104
The key challenge of knowledge distillation is to extract general, moderate and sufficient knowledge from a teacher network to guide a student network. In this paper, a novel Instance Relationship Graph (IRG) is proposed for knowledge distillation. It models three kinds of knowledge, including instance features, instance relationships and feature space transformation, while the latter two kinds of knowledge are neglected by previous methods. Firstly, the IRG is constructed to model the distilled knowledge of one network layer, by considering instance features and instance relationships as vertexes and edges respectively. Secondly, an IRG transformation is proposed to models the feature space transformation across layers. It is more moderate than directly mimicking the features at intermediate layers. Finally, hint loss functions are designed to force a student's IRGs to mimic the structures of a teacher's IRGs. The proposed method effectively captures the knowledge along the whole network via IRGs, and thus shows stable convergence and strong robustness to different network architectures. In addition, the proposed method shows superior performance over existing methods on datasets of various scales.
****

PPGNet: Learning Point-Pair Graph for Line Segment Detection
Ziheng Zhang, Zhengxin Li, Ning Bi, Jia Zheng, Jinlei Wang, Kun Huang, Wei

xin Luo, Yanyu Xu, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7105-7114
In this paper, we present a novel framework to detect line segments in man-made environments. Specifically, we propose to describe junctions, line segments and relationships between them with a simple graph, which is more structured and informative than end-point representation used in existing line segment detection methods. In order to extract a line segment graph from an image, we further introduce the PPGNet, a convolutional neural network that directly infers a graph from an image. We evaluate our method on published benchmarks including York Urban and Wireframe datasets. The results demonstrate that our method achieves satisfactory performance and generalizes well on all the benchmarks. The source code of our work is available at https://github.com/svip-lab/PPGNet.
****

Building Detail-Sensitive Semantic Segmentation Networks With Polynomial Pooling
Zhen Wei, Jingyi Zhang, Li Liu, Fan Zhu, Fumin Shen, Yi Zhou, Si Liu, Yao Sun, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7115-7123
Semantic segmentation is an important computer vision task, which aims to allocate a semantic label to each pixel in an image. When training a segmentation model, it is common to fine-tune a classification network pre-trained on a large-scale dataset. However, as an intrinsic property of the classification model, invariance to spatial perturbation resulting from the lose of detail-sensitivity prevents segmentation networks from achieving high performance. The use of standard poolings is one of the key factors for this invariance. The most common standard poolings are max and average pooling. Max pooling can increase both the invariance to spatial perturbations and the non-linearity of the networks. Average pooling, on the other hand, is sensitive to spatial perturbations, but is a linear function. For semantic segmentation, we prefer both the preservation of detailed cues within a local feature region and non-linearity that increases a network's functional complexity. In this work, we propose a polynomial pooling (P-pooling) function that finds an intermediate form between max and average pooling to provide an optimally balanced and self-adjusted pooling strategy for semantic segmentation. The P-pooling is differentiable and can be applied into a variety of pre-trained networks. Extensive studies on the PASCAL VOC, Cityscapes and ADE20k datasets demonstrate the superiority of P-pooling over other poolings. Experiments on various network architectures and state-of-the-art training strategies also show that models with P-pooling layers consistently outperform those directly fine-tuned using pre-trained classification models.
****

Variational Bayesian Dropout With a Hierarchical Prior
Yuhang Liu, Wenyong Dong, Lei Zhang, Dong Gong, Qinfeng Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7124-7133
Variational dropout (VD) is a generalization of Gaussian dropout, which aims at inferring the posterior of network weights based on a log-uniform prior on them to learn these weights as well as dropout rate simultaneously. The log-uniform prior not only interprets the regularization capacity of Gaussian dropout in network training, but also underpins the inference of such posterior. However, the log-uniform prior is an improper prior (i.e., its integral is infinite), which causes the inference of posterior to be ill-posed, thus restricting the regularization performance of VD. To address this problem, we present a new generalization of Gaussian dropout, termed variational Bayesian dropout (VBD), which turns to exploit a hierarchical prior on the network weights and infer a new joint posterior. Specifically, we implement the hierarchical prior as a zero-mean Gaussian distribution with variance sampled from a uniform hyper-prior. Then, we incorporate such a prior into inferring the joint posterior over network weights and the variance in the hierarchical prior, with which both the network training and dropout rate estimation can be cast into a joint optimization problem. More importantly, the hierarchical prior is a proper prior which enables the inference of posterior to be well-posed. In addition, we further show that the proposed VBD can

be seamlessly applied to network compression. Experiments on classification and network compression demonstrate the superior performance of the proposed VBD in regularizing network training.

****

AANet: Attribute Attention Network for Person Re-Identifications

Chiat-Pin Tay, Sharmili Roy, Kim-Hui Yap; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7134-7143

This paper proposes Attribute Attention Network (AANet), a new architecture that integrates person attributes and attribute attention maps into a classification framework to solve the person re-identification (re-ID) problem. Many person re-ID models typically employ semantic cues such as body parts or human pose to improve the re-ID performance. Attribute information, however, is often not utilized. The proposed AANet leverages on a baseline model that uses body parts and integrates the key attribute information in an unified learning framework. The AANet consists of a global person ID task, a part detection task and a crucial attribute detection task. By estimating the class responses of individual attributes and combining them to form the attribute attention map (AAM), a very strong discriminatory representation is constructed. The proposed AANet outperforms the best state-of-the-art method [??] using ResNet-50 by 3.36% in mAP and 3.12% in Rank-1 accuracy on DukeMTMC-reID dataset. On Market1501 dataset, AANet achieves 92.38% mAP and 95.10% Rank-1 accuracy with re-ranking, outperforming [??], another state of the art method using ResNet-152, by 1.42% in mAP and 0.47% in Rank-1 accuracy. In addition, AANet can perform person attribute prediction (e.g., gender, hair length, clothing length etc.), and localize the attributes in the query image.

****

Overcoming Limitations of Mixture Density Networks: A Sampling and Fitting Framework for Multimodal Future Prediction

Osama Makansi, Eddy Ilg, Ozgun Cicek, Thomas Brox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7144-7153

Future prediction is a fundamental principle of intelligence that helps plan actions and avoid possible dangers. As the future is uncertain to a large extent, modeling the uncertainty and multimodality of the future states is of great relevance. Existing approaches are rather limited in this regard and mostly yield a single hypothesis of the future or, at the best, strongly constrained mixture components that suffer from instabilities in training and mode collapse. In this work, we present an approach that involves the prediction of several samples of the future with a winner-takes-all loss and iterative grouping of samples to multiple modes. Moreover, we discuss how to evaluate predicted multimodal distributions, including the common real scenario, where only a single sample from the ground-truth distribution is available for evaluation. We show on synthetic and real data that the proposed approach triggers good estimates of multimodal distributions and avoids mode collapse.

****

A Main/Subsidiary Network Framework for Simplifying Binary Neural Networks

Yinghao Xu, Xin Dong, Yudian Li, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7154-7162

To reduce memory footprint and run-time latency, techniques such as neural net-work pruning and binarization have been explored separately. However, it is un-clear how to combine the best of the two worlds to get extremely small and efficient models. In this paper, we, for the first time, define the filter-level pruning problem for binary neural networks, which cannot be solved by simply migrating existing structural pruning methods for full-precision models. A novel learning-based approach is proposed to prune filters in our main/subsidiary network frame-work, where the main network is responsible for learning representative features to optimize the prediction performance, and the subsidiary component works as a filter selector on the main network. To avoid gradient mismatch when training the subsidiary component, we propose a layer-wise and bottom-up scheme. We also provide the theoretical and experimental comparison between our learning-ba

sed and greedy rule-based methods. Finally, we empirically demonstrate the effe ctiveness of our approach applied on several binary models, including binarized NIN, VGG-11, and ResNet-18, on various image classification datasets. For bi-na ry ResNet-18 on ImageNet, we use 78.6% filters but can achieve slightly better t est error 49.87% (50.02%-0.15%) than the original model
****

PointNetLK: Robust & Efficient Point Cloud Registration Using PointNet
Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, Simon Lucey; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2019, pp. 7163-7172
PointNet has revolutionized how we think about representing point clouds. For cl assification and segmentation tasks, the approach and its subsequent variants/ex tensions are considered state-of-the-art. To date, the successful application of PointNet to point cloud registration has remained elusive. In this paper we arg ue that PointNet itself can be thought of as a learnable "imaging" function. As a consequence, classical vision algorithms for image alignment can be brought to bear on the problem -- namely the Lucas & Kanade (LK) algorithm. Our central in novations stem from: (i) how to modify the LK algorithm to accommodate the Point Net imaging function, and (ii) unrolling PointNet and the LK algorithm into a si ngle trainable recurrent deep neural network. We describe the architecture, and compare its performance against state-of-the-art in several common registration scenarios. The architecture offers some remarkable properties including: general ization across shape categories and computational efficiency -- opening up new p aths of exploration for the application of deep learning to point cloud registra tion. Code and videos are available at https://github.com/hmgoforth/PointNetLK.
****

Few-Shot Adaptive Faster R-CNN
Tao Wang, Xiaopeng Zhang, Li Yuan, Jiashi Feng; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7173-7182
To mitigate the detection performance drop caused by domain shift, we aim to dev elop a novel few-shot adaptation approach that requires only a few target domain images with limited bounding box annotations. To this end, we first observe sev eral significant challenges. First, the target domain data is highly insufficien t, making most existing domain adaptation methods ineffective. Second, object de tection involves simultaneous localization and classification, further complicat ing the model adaptation process. Third, the model suffers from over-adaptation (similar to overfitting when training with a few data example) and instability r isk that may lead to degraded detection performance in the target domain. To add ress these challenges, we first introduce a pairing mechanism over source and ta rget features to alleviate the issue of insufficient target domain samples. We t hen propose a bi-level module to adapt the source trained detector to the target domain: 1) the split pooling based image level adaptation module uniformly extr acts and aligns paired local patch features over locations, with different scale and aspect ratio; 2) the instance level adaptation module semantically aligns p aired object features while avoids inter-class confusion. Meanwhile, a source mo del feature regularization (SMFR) is applied to stabilize the adaptation process of the two modules. Combining these contributions gives a novel few-shot adapti ve Faster-RCNN framework, termed FAFRCNN, which effectively adapts to target dom ain with a few labeled samples. Experiments with multiple datasets show that our model achieves new state-of-the-art performance under both the interested few-s hot domain adaptation(FDA) and unsupervised domain adaptation(UDA) setting.
****

VRSTC: Occlusion-Free Video Person Re-Identification
Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2019, pp. 7183-7192
Video person re-identification (re-ID) plays an important role in surveillance v ideo analysis. However, the performance of video re-ID degenerates severely unde r partial occlusion. In this paper, we propose a novel network, called Spatio-Te mporal Completion network (STCnet), to explicitly handle partial occlusion probl

em. Different from most previous works that discard the occluded frames, STCnet can recover the appearance of the occluded parts. For one thing, the spatial structure of a pedestrian frame can be used to predict the occluded body parts from the unoccluded body parts of this frame. For another, the temporal patterns of pedestrian sequence provide important clues to generate the contents of occluded parts. With the spatio-temporal information, STCnet can recover the appearance for the occluded parts, which could be leveraged with those unoccluded parts for more accurate video re-ID. By combining a re-ID network with STCnet, a video re-ID framework robust to partial occlusion (VRSTC) is proposed. Experiments on three challenging video re-ID databases demonstrate that the proposed approach outperforms the state-of-the-arts.

****

Compact Feature Learning for Multi-Domain Image Classification

Yajing Liu, Xinmei Tian, Ya Li, Zhiwei Xiong, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7193-7201

The goal of multi-domain learning is to improve the performance over multiple domains by making full use of all training data from them. However, variations of feature distributions across different domains result in a non-trivial solution of multi-domain learning. The state-of-the-art work regarding multi-domain classification aims to extract domain-invariant features and domain-specific features independently. However, they view the distributions of features from different classes as a general distribution and try to match these distributions across domains, which lead to the mixture of features from different classes across domains and degrade the performance of classification. Additionally, existing works only force the shared features among domains to be orthogonal to the features in the domain-specific network. However, redundant features between the domain-specific networks still remain, which may shrink the discriminative ability of domain-specific features. Therefore, we propose an end-to-end network to obtain the more optimal features, which we call compact features. We propose to extract the domain-invariant features by matching the joint distributions of different domains, which have dis- tinct boundaries between different classes. Moreover, we add an orthogonal constraint between the private features across domains to ensure the discriminative ability of the domain-specific space. The proposed method is validated on three landmark datasets, and the results demonstrate the effectiveness of our method.

****

Adaptive Transfer Network for Cross-Domain Person Re-Identification

Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7202-7211

Recent deep learning based person re-identification approaches have steadily improved the performance for benchmarks, however they often fail to generalize well from one domain to another. In this work, we propose a novel adaptive transfer network (ATNet) for effective cross-domain person re-identification. ATNet looks into the essential causes of domain gap and addresses it following the principle of "divide-and-conquer". It decomposes the complicated cross-domain transfer into a set of factor-wise sub-transfers, each of which concentrates on style transfer with respect to a certain imaging factor, e.g., illumination, resolution and camera view etc. An adaptive ensemble strategy is proposed to fuse factor-wise transfers by perceiving the affect magnitudes of various factors on images. Such "decomposition-and-ensemble" strategy gives ATNet the capability of precise style transfer at factor level and eventually effective transfer across domains. In particular, ATNet consists of a transfer network composed by multiple factor-wise CycleGANs and an ensemble CycleGAN as well as a selection network that infers the affects of different factors on transferring each image. Extensive experimental results on three widely-used datasets, i.e., Market-1501, DukeMTMC-reID and PRID2011 have demonstrated the effectiveness of the proposed ATNet with significant performance improvements over state-of-the-art methods.

****

Large-Scale Few-Shot Learning: Knowledge Transfer With Class Hierarchy
Aoxue Li, Tiange Luo, Zhiwu Lu, Tao Xiang, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7212-7220
Recently, large-scale few-shot learning (FSL) becomes topical. It is discovered that, for a large-scale FSL problem with 1,000 classes in the source domain, a strong baseline emerges, that is, simply training a deep feature embedding model using the aggregated source classes and performing nearest neighbor (NN) search using the learned features on the target classes. The state-of-the-art large-scale FSL methods struggle to beat this baseline, indicating intrinsic limitations on scalability. To overcome the challenge, we propose a novel large-scale FSL model by learning transferable visual features with the class hierarchy which encodes the semantic relations between source and target classes. Extensive experiments show that the proposed model significantly outperforms not only the NN baseline but also the state-of-the-art alternatives. Furthermore, we show that the proposed model can be easily extended to the large-scale zero-shot learning (ZSL) problem and also achieves the state-of-the-art results.
****
Moving Object Detection Under Discontinuous Change in Illumination Using Tensor Low-Rank and Invariant Sparse Decomposition
Moein Shakeri, Hong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7221-7230
Although low-rank and sparse decomposition based methods have been successfully applied to the problem of moving object detection using structured sparsity-inducing norms, they are still vulnerable to significant illumination changes that arise in certain applications. We are interested in moving object detection in applications involving time-lapse image sequences for which current methods mistakenly group moving objects and illumination changes into foreground. Our method relies on the multilinear (tensor) data low-rank and sparse decomposition framework to address the weaknesses of existing methods. The key to our proposed method is to create first a set of prior maps that can characterize the changes in the image sequence due to illumination. We show that they can be detected by a k-support norm. To deal with concurrent, two types of changes, we employ two regularization terms, one for detecting moving objects and the other for accounting for illumination changes, in the tensor low-rank and sparse decomposition formulation. Through comprehensive experiments using challenging datasets, we show that our method demonstrates a remarkable ability to detect moving objects under discontinuous change in illumination, and outperforms the state-of-the-art solutions to this challenging problem.
****
Pedestrian Detection With Autoregressive Network Phases
Garrick Brazil, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7231-7240
We present an autoregressive pedestrian detection framework with cascaded phases designed to progressively improve precision. The proposed framework utilizes a novel lightweight stackable decoder-encoder module which uses convolutional re-sampling layers to improve features while maintaining efficient memory and runtime cost. Unlike previous cascaded detection systems, our proposed framework is designed within a region proposal network and thus retains greater context of nearby detections compared to independently processed RoI systems. We explicitly encourage increasing levels of precision by assigning strict labeling policies to each consecutive phase such that early phases develop features primarily focused on achieving high recall and later on accurate precision. In consequence, the final feature maps form more peaky radial gradients emulating from the centroids of unique pedestrians. Using our proposed autoregressive framework leads to new state-of-the-art performance on the reasonable and occlusion settings of the Caltech pedestrian dataset, and achieves competitive state-of-the-art performance on the KITTI dataset.
****
All You Need Is a Few Shifts: Designing Efficient Convolutional Neural Networks

for Image Classification
Weijie Chen, Di Xie, Yuan Zhang, Shiliang Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7241-7250
Shift operation is an efficient alternative over depthwise separable convolution. However, it is still bottlenecked by its implementation manner, namely memory movement. To put this direction forward, a new and novel basic component named Sparse Shift Layer (SSL) is introduced in this paper to construct efficient convolutional neural networks. In this family of architectures, the basic block is only composed by 1x1 convolutional layers with only a few shift operations applied to the intermediate feature maps. To make this idea feasible, we introduce shift operation penalty during optimization and further propose a quantization-aware shift learning method to impose the learned displacement more friendly for inference. Extensive ablation studies indicate that only a few shift operations are sufficient to provide spatial information communication. Furthermore, to maximize the role of SSL, we redesign an improved network architecture to Fully Exploit the limited capacity of neural Network (FE-Net). Equipped with SSL, this network can achieve 75.0% top-1 accuracy on ImageNet with only 563M M-Adds. It surpasses other counterparts constructed by depthwise separable convolution and the networks searched by NAS in terms of accuracy and practical speed.
****
Stochastic Class-Based Hard Example Mining for Deep Metric Learning
Yumin Suh, Bohyung Han, Wonsik Kim, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7251-7259
Performance of deep metric learning depends heavily on the capability of mining hard negative examples during training. However, many metric learning algorithms often require intractable computational cost due to frequent feature computations and nearest neighbor searches in a large-scale dataset. As a result, existing approaches often suffer from trade-off between training speed and prediction accuracy. To alleviate this limitation, we propose a stochastic hard negative mining method. Our key idea is to adopt class signatures that keep track of feature embedding online with minor additional cost during training, and identify hard negative example candidates using the signatures. Given an anchor instance, our algorithm first selects a few hard negative classes based on the class-to-sample distances and then performs a refined search in an instance-level only from the selected classes. As most of the classes are discarded at the first step, it is much more efficient than exhaustive search while effectively mining a large number of hard examples. Our experiment shows that the proposed technique improves image retrieval accuracy substantially; it achieves the state-of-the-art performance on the several standard benchmark datasets.
****
Revisiting Local Descriptor Based Image-To-Class Measure for Few-Shot Learning
Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7260-7268
Few-shot learning in image classification aims to learn a classifier to classify images when only few training examples are available for each class. Recent work has achieved promising classification performance, where an image-level feature based measure is usually used. In this paper, we argue that a measure at such a level may not be effective enough in light of the scarcity of examples in few-shot learning. Instead, we think a local descriptor based image-to-class measure should be taken, inspired by its surprising success in the heydays of local invariant features. Specifically, building upon the recent episodic training mechanism, we propose a Deep Nearest Neighbor Neural Network (DN4 in short) and train it in an end-to-end manner. Its key difference from the literature is the replacement of the image-level feature based measure in the final layer by a local descriptor based image-to-class measure. This measure is conducted online via a k-nearest neighbor search over the deep local descriptors of convolutional feature maps. The proposed DN4 not only learns the optimal deep local descriptors for the image-to-class measure, but also utilizes the higher efficiency of such a meas

ure in the case of example scarcity, thanks to the exchangeability of visual patterns across the images in the same class. Our work leads to a simple, effective, and computationally efficient framework for few-shot learning. Experimental study on benchmark datasets consistently shows its superiority over the related state-of-the-art, with the largest absolute improvement of 17% over the next best. The source code can be available from https://github.com/WenbinLee/DN4.git.
****

Towards Robust Curve Text Detection With Conditional Spatial Expansion
Zichuan Liu, Guosheng Lin, Sheng Yang, Fayao Liu, Weisi Lin, Wang Ling Goh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7269-7278
It is challenging to detect curve texts due to their irregular shapes and varying sizes. In this paper, we first investigate the deficiency of the existing curve detection methods and then propose a novel Conditional Spatial Expansion (CSE) mechanism to improve the performance of curve detection. Instead of regarding the curve text detection as a polygon regression or a segmentation problem, we formulate it as a sequence prediction on the spatial domain. CSE starts with a seed arbitrarily chosen within a text region and progressively merges neighborhood regions based on the extracted local features by a CNN and contextual information of merged regions. The CSE is highly parameterized and can be seamlessly integrated into existing object detection frameworks. Enhanced by the data-dependent CSE mechanism, our curve text detection system provides robust instance-level text region extraction with minimal post-processing. The analysis experiment shows that our CSE can handle texts with various shapes, sizes, and orientations, and can effectively suppress the false-positives coming from text-like textures or unexpected texts included in the same RoI. Compared with the existing curve text detection algorithms, our method is more robust and enjoys a simpler processing flow. It also creates a new state-of-art performance on curve text benchmarks with F-measurement of up to 78.4%.
****

Revisiting Perspective Information for Efficient Crowd Counting
Miaojing Shi, Zhaohui Yang, Chao Xu, Qijun Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7279-7288
Crowd counting is the task of estimating people numbers in crowd images. Modern crowd counting methods employ deep neural networks to estimate crowd counts via crowd density regressions. A major challenge of this task lies in the perspective distortion, which results in drastic person scale change in an image. Density regression on the small person area is in general very hard. In this work, we propose a perspective-aware convolutional neural network (PACNN) for efficient crowd counting, which integrates the perspective information into density regression to provide additional knowledge of the person scale change in an image. Ground truth perspective maps are firstly generated for training; PACNN is then specifically designed to predict multi-scale perspective maps and encode them as perspective-aware weighting layers in the network to adaptively combine the outputs of multi-scale density maps. The weights are learned at every pixel of the maps such that the final density combination is robust to the perspective distortion. We conduct extensive experiments on the ShanghaiTech, WorldExpo'10, UCF_CC_50, and UCSD datasets, and demonstrate the effectiveness and efficiency of PACNN over the state-of-the-art.
****

Towards Universal Object Detection by Domain Attention
Xudong Wang, Zhaowei Cai, Dashan Gao, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7289-7298
Despite increasing efforts on universal representations for visual recognition, few have addressed object detection. In this paper, we develop an effective and efficient universal object detection system that is capable of working on various image domains, from human faces and traffic signs to medical CT images. Unlike multi-domain models, this universal model does not require prior knowledge of t

he domain of interest. This is achieved by the introduction of a new family of a daptation layers, based on the principles of squeeze and excitation, and a new d omain-attention mechanism. In the proposed universal detector, all parameters an d computations are shared across domains, and a single network processes all dom ains all the time. Experiments, on a newly established universal object detectio n benchmark of 11 diverse datasets, show that the proposed detector outperforms a bank of individual detectors, a multi-domain detector, and a baseline universa l detector, with a 1.3x parameter increase over a single-domain baseline detecto r. The code and benchmark are available at http://www.svcl.ucsd.edu/projects/uni versal-detection/.

****

Ensemble Deep Manifold Similarity Learning Using Hard Proxies

Nicolas Aziere, Sinisa Todorovic; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2019, pp. 7299-7307

This paper is about learning deep representations of images such that images bel onging to the same class have more similar representations than those belonging to different classes. For this goal, prior work typically uses the triplet or N-pair loss, specified in terms of either l2-distances or dot-products between dee p features. However, such formulations seem poorly suited to the highly non-Euc lidean deep feature space. Our first contribution is in specifying the N-pair lo ss in terms of manifold similarities between deep features. We introduce a new t ime- and memory-efficient method for estimating the manifold similarities by usi ng a closed-form convergence solution of the Random Walk algorithm. Our efficien cy comes, in part, from following the recent work that randomly partitions the d eep feature space, and expresses image distances via representatives of the resu lting subspaces, a.k.a. proxies. Our second contribution is aimed at reducing ov erfitting by estimating hard proxies that are as close to one another as possibl e, but remain in their respective subspaces. Our evaluation demonstrates that we outperform the state of the art in both image retrieval and clustering on the b enchmark CUB-200-2011, Cars196, and Stanford Online Products datasets.

****

Quantization Networks

Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqi ang Huang, Xian-sheng Hua; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2019, pp. 7308-7316

Although deep neural networks are highly effective, their high computational and memory costs severely hinder their applications to portable devices. As a conse quence, lowbit quantization, which converts a full-precision neural network into a low-bitwidth integer version, has been an active and promising research topic . Existing methods formulate the low-bit quantization of networks as an approxim ation or optimization problem. Approximation-based methods confront the gradient mismatch problem, while optimizationbased methods are only suitable for quantiz ing weights and can introduce high computational cost during the training stage. In this paper, we provide a simple and uniform way for weights and activations quantization by formulating it as a differentiable non-linear function. The quan tization function is represented as a linear combination of several Sigmoid func tions with learnable biases and scales that could be learned in a lossless and e nd-to-end manner via continuous relaxation of the steepness of Sigmoid functions . Extensive experiments on image classification and object detection tasks show that our quantization networks outperform state-of-the-art methods. We believe t hat the proposed method will shed new lights on the interpretation of neural net work quantization.

****

RES-PCA: A Scalable Approach to Recovering Low-Rank Matrices

Chong Peng, Chenglizhao Chen, Zhao Kang, Jianbo Li, Qiang Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 019, pp. 7317-7325

Robust principal component analysis (RPCA) has drawn significant attentions due to its powerful capability in recovering low-rank matrices as well as successful appplications in various real world problems. The current state-of-the-art algo

rithms usually need to solve singular value decomposition of large matrices, whi ch generally has at least a quadratic or even cubic complexity. This drawback ha s limited the application of RPCA in solving real world problems. To combat this drawback, in this paper we propose a new type of RPCA method, RES-PCA, which is linearly efficient and scalable in both data size and dimension. For comparison purpose, AltProj, an existing scalable approach to RPCA requires the precise kn owlwdge of the true rank; otherwise, it may fail to recover low-rank matrices. B y contrast, our method works with or without knowing the true rank; even when bo th methods work, our method is faster. Extensive experiments have been performed and testified to the effectiveness of proposed method quantitatively and in vis ual quality, which suggests that our method is suitable to be employed as a ligh t-weight, scalable component for RPCA in any application pipelines.

****

Occlusion-Net: 2D/3D Occluded Keypoint Localization Using Graph Networks

N. Dinesh Reddy, Minh Vo, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7326-73 35

We present Occlusion-Net, a framework to predict 2D and 3D locations of occluded keypoints for objects, in a largely self-supervised manner. We use an off-the-s helf detector as input (like MaskRCNN) that is trained only on visible key point annotations. This is the only supervision used in this work. A graph encoder ne twork then explicitly classifies invisible edges and a graph decoder network cor rects the occluded keypoint locations from the initial detector. Central to this work is a trifocal tensor loss that provides indirect self-supervision for occl uded keypoint locations that are visible in other views of the object. The 2D ke ypoints are then passed into a 3D graph network that estimates the 3D shape and camera pose using the self-supervised re-projection loss. At test time, our appr oach successfully localizes keypoints in a single view under a diverse set of se vere occlusion settings. We demonstrate and evaluate our approach on synthetic C AD data as well as a large image set capturing vehicles at many busy city inters ections. As an interesting aside, we compare the accuracy of human labels of inv isible keypoints against those obtained from geometric trifocal-tensor loss.

****

Efficient Featurized Image Pyramid Network for Single Shot Detector

Yanwei Pang, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, Ling Shao ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2019, pp. 7336-7344

Single-stage object detectors have recently gained popularity due to their combi ned advantage of high detection accuracy and real-time speed. However, while pro mising results have been achieved by these detectors on standard-sized objects, their performance on small objects is far from satisfactory. To detect very smal l/large objects, classical pyramid representation can be exploited, where an ima ge pyramid is used to build a feature pyramid (featurized image pyramid), enabli ng detection across a range of scales. Existing single-stage detectors avoid suc h a featurized image pyramid representation due to its memory and time complexit y. In this paper, we introduce a light-weight architecture to efficiently produc e featurized image pyramid in a single-stage detection framework. The resulting multi-scale features are then injected into the prediction layers of the detecto r using an attention module. The performance of our detector is validated on two benchmarks: PASCAL VOC and MS COCO. For a 300x300 input, our detector operates at 111 frames per second (FPS) on a Titan X GPU, providing state-of-the-art dete ction accuracy on PASCAL VOC 2007 testset. On the MS COCO testset, our detector achieves state-of-the-art results surpassing all existing single-stage methods i n the case of single-scale inference.

****

Multi-Task Multi-Sensor Fusion for 3D Object Detection

Ming Liang, Bin Yang, Yun Chen, Rui Hu, Raquel Urtasun; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7345-7353

In this paper we propose to exploit multiple related tasks for accurate multi-s

ensor 3D object detection. Towards this goal we present an end-to-end learnable architecture that reasons about 2D and 3D object detection as well as ground estimation and depth completion. Our experiments show that all these tasks are complementary and help the network learn better representations by fusing information at various levels. Importantly, our approach leads the KITTI benchmark on 2D, 3D and bird's eye view object detection, while being real-time.
****

Domain-Specific Batch Normalization for Unsupervised Domain Adaptation
Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7354-7362
We propose a novel unsupervised domain adaptation framework based on domain-specific batch normalization in deep neural networks. We aim to adapt to both domains by specializing batch normalization layers in convolutional neural networks while allowing them to share all other model parameters, which is realized by a two-stage algorithm. In the first stage, we estimate pseudo-labels for the examples in the target domain using an external unsupervised domain adaptation algorithm---for example, MSTN or CPUA---integrating the proposed domain-specific batch normalization. The second stage learns the final models using a multi-task classification loss for the source and target domains. Note that the two domains have separate batch normalization layers in both stages. Our framework can be easily incorporated into the domain adaptation techniques based on deep neural networks with batch normalization layers. We also present that our approach can be extended to the problem with multiple source domains. The proposed algorithm is evaluated on multiple benchmark datasets and achieves the state-of-the-art accuracy in the standard setting and the multi-source domain adaption scenario.
****

Grid R-CNN
Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, Junjie Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7363-7372
This paper proposes a novel object detection framework named Grid R-CNN, which adopts a grid guided localization mechanism for accurate object detection. Different from the traditional regression based methods, the Grid R-CNN captures the spatial information explicitly and enjoys the position sensitive property of fully convolutional architecture. Instead of using only two independent points, we design a multi-point supervision formulation to encode more clues in order to reduce the impact of inaccurate prediction of specific points. To take the full advantage of the correlation of points in a grid, we propose a two-stage information fusion strategy to fuse feature maps of neighbor grid points. The grid guided localization approach is easy to be extended to different state-of-the-art detection frameworks. Grid R-CNN leads to high quality object localization, and experiments demonstrate that it achieves a 4.1% AP gain at IoU=0.8 and a 10.0% AP gain at IoU=0.9 on COCO benchmark compared to Faster R-CNN with Res50 backbone and FPN architecture.
****

MetaCleaner: Learning to Hallucinate Clean Representations for Noisy-Labeled Visual Recognition
Weihe Zhang, Yali Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7373-7382
Deep Neural Networks (DNNs) have achieved remarkable successes in large-scale visual recognition. However, they often suffer from overfitting under noisy labels. To alleviate this problem, we propose a conceptually simple but effective MetaCleaner, which can learn to hallucinate a clean representation of an object category, according to a small noisy subset from the same category. Specially, MetaCleaner consists of two flexible submodules. The first submodule, namely Noisy Weighting, can estimate the confidence scores of all the images in the noisy subset, by analyzing their deep features jointly. The second submodule, namely Clean Hallucinating, can generate a clean representation from the noisy subset, by summarizing the noisy images with their confidence scores. Via MetaCleaner, DNNs ca

n strengthen its robustness to noisy labels, as well as enhance its generalizati on capacity with richer data diversity. Moreover, MetaCleaner can be easily inte grated into the standard training procedure of DNNs, which promotes its value fo r real-life applications. We conduct extensive experiments on two popular benchm arks in noisy-labeled recognition, i.e., Food-101N and Clothing1M. For both data sets, our MetaCleaner significantly outperforms baselines, and achieves the stat e-of-the-art performance.
****

Mapping, Localization and Path Planning for Image-Based Navigation Using Visual Features and Map

Janine Thoma, Danda Pani Paudel, Ajad Chhatkuli, Thomas Probst, Luc Van Gool ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2019, pp. 7383-7391

Building on progress in feature representations for image retrieval, image-based localization has seen a surge of research interest. Image-based localization ha s the advantage of being inexpensive and efficient, often avoiding the use of 3D metric maps altogether. That said, the need to maintain a large amount of refer ence images as an effective support of localization in a scene, nonetheless call s for them to be organized in a map structure of some kind. The problem of local ization often arises as part of a navigation process. We are, therefore, interes ted in summarizing the reference images as a set of landmarks, which meet the re quirements for image-based navigation. A contribution of this paper is to formul ate such a set of requirements for the two sub-tasks involved: compact map const ruction and accurate self localization. These requirements are then exploited fo r compact map representation and accurate self-localization, using the framework of a network flow problem. During this process, we formulate the map constructi on and self-localization problems as convex quadratic and second-order cone prog rams, respectively. We evaluate our methods on publicly available indoor and out door datasets, where they outperform existing methods significantly.
****

Triply Supervised Decoder Networks for Joint Detection and Segmentation

Jiale Cao, Yanwei Pang, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7392-7401

Joint object detection and semantic segmentation is essential in many fields suc h as self-driving cars. An initial attempt towards this goal is to simply share a single network for multi-task learning. We argue that it does not make full us e of the fact that detection and segmentation are mutually beneficial. In this p aper, we propose a framework called TripleNet to deeply boost these two tasks. O n the one hand, to deeply join the two tasks at different scales, triple supervi sions including detection-oriented supervision and class-aware/agnostic segmenta tion supervisions are imposed on each layer of the decoder. Class-agnostic segme ntation provides an objectness prior to detection and segmentation. On the other hand, to further intercross the two tasks and refine the features in each scale , two light-weight modules (i.e., the inner-connected module and the attention s kip-layer fusion) are incorporated. Because segmentation supervision on each dec oder layer are not performed at the test stage and two added modules are light-w eight, the proposed TripleNet can run at a real-time speed (16 fps). Experiments on the VOC 2007/2012 and COCO datasets show that TripleNet outperforms all the other one-stage methods on both two tasks (e.g., 81.9% mAP and 83.3% mIoU on VOC 2012, and 37.1% mAP and 59.6% mIoU on COCO) by a single network.
****

Leveraging the Invariant Side of Generative Zero-Shot Learning

Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, Zi Huang; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2019, pp. 7402-7411

Conventional zero-shot learning (ZSL) methods generally learn an embedding, e.g. , visual-semantic mapping, to handle the unseen visual samples via an indirect m anner. In this paper, we take the advantage of generative adversarial networks (GANs) and propose a novel method, named leveraging invariant side GAN (LisGAN), which can directly generate the unseen features from random noises which are con

ditioned by the semantic descriptions. Specifically, we train a conditional Wasserstein GANs in which the generator synthesizes fake unseen features from noises and the discriminator distinguishes the fake from real via a minimax game. Considering that one semantic description can correspond to various synthesized visual samples, and the semantic description, figuratively, is the soul of the generated features, we introduce soul samples as the invariant side of generative zero-shot learning in this paper. A soul sample is the meta-representation of one class. It visualizes the most semantically-meaningful aspects of each sample in the same category. We regularize that each generated sample (the varying side of generative ZSL) should be close to at least one soul sample (the invariant side) which has the same class label with it. At the zero-shot recognition stage, we propose to use two classifiers, which are deployed in a cascade way, to achieve a coarse-to-fine result. Experiments on five popular benchmarks verify that our proposed approach can outperform state-of-the-art methods with significant improvements.

****

Exploring the Bounds of the Utility of Context for Object Detection

Ehud Barnea, Ohad Ben-Shahar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7412-7420

The recurring context in which objects appear holds valuable information that can be employed to predict their existence. This intuitive observation indeed led many researchers to endow appearance-based detectors with explicit reasoning about context. The underlying thesis suggests that stronger contextual relations would facilitate greater improvements in detection capacity. In practice, however, the observed improvement in many case is modest at best, and often only marginal. In this work we seek to improve our understanding of this phenomenon, in part by pursuing an opposite approach. Instead of attempting to improve detection scores by employing context, we treat the utility of context as an optimization problem: to what extent can detection scores be improved by considering context or any other kind of additional information? With this approach we explore the bounds on improvement by using contextual relations between objects and provide a tool for identifying the most helpful ones. We show that simple co-occurrence relations can often provide large gains, while in other cases a significant improvement is simply impossible or impractical with either co-occurrence or more precise spatial relations. To better understand these results we then analyze the ability of context to handle different types of false detections, revealing that tested contextual information cannot ameliorate localization errors, severely limiting its gains. These and additional insights further our understanding on where and why utilization of context for object detection succeeds and fails.

****

A-CNN: Annularly Convolutional Neural Networks on Point Clouds

Artem Komarichev, Zichun Zhong, Jing Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7421-7430

Analyzing the geometric and semantic properties of 3D point clouds through the deep networks is still challenging due to the irregularity and sparsity of samplings of their geometric structures. This paper presents a new method to define and compute convolution directly on 3D point clouds by the proposed annular convolution. This new convolution operator can better capture the local neighborhood geometry of each point by specifying the (regular and dilated) ring-shaped structures and directions in the computation. It can adapt to the geometric variability and scalability at the signal processing level. We apply it to the developed hierarchical neural networks for object classification, part segmentation, and semantic segmentation in large-scale scenes. The extensive experiments and comparisons demonstrate that our approach outperforms the state-of-the-art methods on a variety of standard benchmark datasets (e.g., ModelNet10, ModelNet40, ShapeNet-part, S3DIS, and ScanNet).

****

DARNet: Deep Active Ray Network for Building Segmentation

Dominic Cheng, Renjie Liao, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.

7431-7439
In this paper, we propose a Deep Active Ray Network (DARNet) for automatic building segmentation. Taking an image as input, it first exploits a deep convolutional neural network (CNN) as the backbone to predict energy maps, which are further utilized to construct an energy function. A polygon-based contour is then evolved via minimizing the energy function, of which the minimum defines the final segmentation. Instead of parameterizing the contour using Euclidean coordinates, we adopt polar coordinates, i.e., rays, which not only prevents self-intersection but also simplifies the design of the energy function. Moreover, we propose a loss function that directly encourages the contours to match building boundaries. Our DARNet is trained end-to-end by back-propagating through the energy minimization and the backbone CNN, which makes the CNN adapt to the dynamics of the contour evolution. Experiments on three building instance segmentation datasets demonstrate our DARNet achieves either state-of-the-art or comparable performances to other competitors.
****

Point Cloud Oversegmentation With Graph-Structured Deep Metric Learning
Loic Landrieu, Mohamed Boussaha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7440-7449
We propose a new supervized learning framework for oversegmenting 3D point clouds into superpoints. We cast this problem as learning deep embeddings of the local geometry and radiometry of 3D points, such that the border of objects presents high contrasts. The embeddings are computed using a lightweight neural network operating on the points' local neighborhood. Finally, we formulate point cloud oversegmentation as a graph partition problem with respect to the learned embeddings. This new approach allows us to set a new state-of-the-art in point cloud oversegmentation by a significant margin, on a dense indoor dataset (S3DIS) and a sparse outdoor one (vKITTI). Our best solution requires over five times fewer superpoints to reach similar performance than previously published methods on S3DIS. Furthermore, we show that our framework can be used to improve superpoint-based semantic segmentation algorithms, setting a new state-of-the-art for this task as well.
****

Graphonomy: Universal Human Parsing via Graph Transfer Learning
Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7450-7459
Prior highly-tuned human parsing models tend to fit towards each dataset in a specific domain or with discrepant label granularity, and can hardly be adapted to other human parsing tasks without extensive re-training. In this paper, we aim to learn a single universal human parsing model that can tackle all kinds of human parsing needs by unifying label annotations from different domains or at various levels of granularity. This poses many fundamental learning challenges, e.g. discovering underlying semantic structures among different label granularity, performing proper transfer learning across different image domains, and identifying and utilizing label redundancies across related tasks. To address these challenges, we propose a new universal human parsing agent, named "Graphonomy", which incorporates hierarchical graph transfer learning upon the conventional parsing network to encode the underlying label semantic structures and propagate relevant semantic information. In particular, Graphonomy first learns and propagates compact high-level graph representation among the labels within one dataset via Intra-Graph Reasoning, and then transfers semantic information across multiple datasets via Inter-Graph Transfer. Various graph transfer dependencies (e.g., similarity, linguistic knowledge) between different datasets are analyzed and encoded to enhance graph transfer capability. By distilling universal semantic graph representation to each specific task, Graphonomy is able to predict all levels of parsing labels in one system without piling up the complexity. Experimental results show Graphonomy effectively achieves the state-of-the-art results on three human parsing benchmarks as well as advantageous universal human parsing performance.

****
Fitting Multiple Heterogeneous Models by Multi-Class Cascaded T-Linkage
Luca Magri, Andrea Fusiello; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7460-7468
This paper addresses the problem of multiple models fitting in the general context where the sought structures can be described by a mixture of heterogeneous parametric models drawn from different classes. To this end, we conceive a multi-model selection framework that extend T-linkage to cope with different nested class of models. Our method, called MCT, compares favourably with the state-of-the-art on publicly available data-sets for various fitting problems: lines and conics, homographies and fundamental matrices, planes and cylinders.
****
A Late Fusion CNN for Digital Matting
Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, Weiwei Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7469-7478
This paper studies the structure of a deep convolutional neural network to predict the foreground alpha matte by taking a single RGB image as input. Our network is fully convolutional with two decoder branches for the foreground and background classification respectively. Then a fusion branch is used to integrate the two classification results which gives rise to alpha values as the soft segmentation result. This design provides more degrees of freedom than a single decoder branch for the network to obtain better alpha values during training. The network can implicitly produce trimaps without user interaction, which is easy to use for novices without expertise in digital matting. Experimental results demonstrate that our network can achieve high-quality alpha mattes for various types of objects and outperform the state-of-the-art CNN-based image matting methods on the human image matting task.
****
BASNet: Boundary-Aware Salient Object Detection
Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, Martin Jagersand; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7479-7489
Deep Convolutional Neural Networks have been adopted for salient object detection and achieved the state-of-the-art performance. Most of the previous works however focus on region accuracy but not on the boundary quality. In this paper, we propose a predict-refine architecture, BASNet, and a new hybrid loss for Boundary-Aware Salient object detection. Specifically, the architecture is composed of a densely supervised Encoder-Decoder network and a residual refinement module, which are respectively in charge of saliency prediction and saliency map refinement. The hybrid loss guides the network to learn the transformation between the input image and the ground truth in a three-level hierarchy -- pixel-, patch- and map- level -- by fusing Binary Cross Entropy (BCE), Structural SIMilarity (SSIM) and Intersection-over-Union (IoU) losses. Equipped with the hybrid loss, the proposed predict-refine architecture is able to effectively segment the salient object regions and accurately predict the fine structures with clear boundaries. Experimental results on six public datasets show that our method outperforms the state-of-the-art methods both in terms of regional and boundary evaluation measures. Our method runs at over 25 fps on a single GPU. The code is available at: https://github.com/NathanUA/BASNet.
****
ZigZagNet: Fusing Top-Down and Bottom-Up Context for Object Segmentation
Di Lin, Dingguo Shen, Siting Shen, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, Hui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7490-7499
Multi-scale context information has proven to be essential for object segmentation tasks. Recent works construct the multi-scale context by aggregating convolutional feature maps extracted by different levels of a deep neural network. This is typically done by propagating and fusing features in a one-directional, top-down and bottom-up, manner. In this work, we introduce ZigZagNet, which aggregate

s a richer multi-context feature map by using not only dense top-down and bottom-up propagation, but also by introducing pathways crossing between different levels of the top-down and the bottom-up hierarchies, in a zig-zag fashion. Furthermore, the context information is exchanged and aggregated over multiple stages, where the fused feature maps from one stage are fed into the next one, yielding a more comprehensive context for improved segmentation performance. Our extensive evaluation on the public benchmarks demonstrates that ZigZagNet surpasses the state-of-the-art accuracy for both semantic segmentation and instance segmentation tasks.
****

Object Instance Annotation With Deep Extreme Level Set Evolution
Zian Wang, David Acuna, Huan Ling, Amlan Kar, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7500-7508
In this paper, we tackle the task of interactive object segmentation. We revive the old ideas on level set segmentation which framed object annotation as curve evolution. Carefully designed energy functions ensured that the curve was well aligned with image boundaries, and generally "well behaved". The Level Set Method can handle objects with complex shapes and topological changes such as merging and splitting, thus able to deal with occluded objects and objects with holes. We propose Deep Extreme Level Set Evolution that combines powerful CNN models with level set optimization in an end-to-end fashion. Our method learns to predict evolution parameters conditioned on the image and evolves the predicted initial contour to produce the final result. We make our model interactive by incorporating user clicks on the extreme boundary points, following DEXTR. We show that our approach significantly outperforms DEXTR on the static Cityscapes dataset and the video segmentation benchmark DAVIS, and performs on par on PASCAL and SBD.
****

Leveraging Crowdsourced GPS Data for Road Extraction From Aerial Imagery
Tao Sun, Zonglin Di, Pengyu Che, Chun Liu, Yin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7509-7518
Deep learning is revolutionizing the mapping industry. Under lightweight human curation, computer has generated almost half of the roads in Thailand on Open- StreetMap (OSM) using high resolution aerial imagery. Bing maps are displaying 125 million computer generated building polygons in the U.S. While tremendously more efficient than manual mapping, one cannot map out everything from the air. Especially for roads, a small prediction gap by image occlusion renders the entire road useless for routing. Misconnections can be more dangerous. Therefore computer-based mapping often requires local verifications, which is still labor intensive. In this paper, we propose to leverage crowdsourced GPS data to improve and support road extraction from aerial imagery. Through novel data augmentation, GPS rendering, and 1D transpose convolution techniques, we show almost 5% improvements over previous competition winning models, and much better robustness when predicting new areas without any new training data or domain adaptation.
****

Adaptive Pyramid Context Network for Semantic Segmentation
Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7519-7528
Recent studies witnessed that context features can significantly improve the performance of deep semantic segmentation networks. Current context based segmentation methods differ with each other in how to construct context features and perform differently in practice. This paper firstly introduces three desirable properties of context features in segmentation task. Specially, we find that Global-guided Local Affinity (GLA) can play a vital role in constructing effective context features, while this property has been largely ignored in previous works. Based on this analysis, this paper proposes Adaptive Pyramid Context Network (APCNet) for semantic segmentation. APCNet adaptively constructs multi-scale contextual representations with multiple well-designed Adaptive Context Modules (ACMs). S

pecifically, each ACM leverages a global image representation as a guidance to e stimate the local affinity coefficients for each sub-region, and then calculates a context vector with these affinities. We empirically evaluate our APCNet on t hree semantic segmentation and scene parsing datasets, including PASCAL VOC 2012 , Pascal-Context, and ADE20K dataset. Experimental results show that APCNet ach ieves state-of-the-art performance on all three benchmarks, and obtains a new re cord 84.2% on PASCAL VOC 2012 test set without MS COCO pre-trained and any post- processing.

****

Isospectralization, or How to Hear Shape, Style, and Correspondence

Luca Cosmo, Mikhail Panine, Arianna Rampini, Maks Ovsjanikov, Michael M. Bro nstein, Emanuele Rodola; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2019, pp. 7529-7538

The question whether one can recover the shape of a geometric object from its La placian spectrum ('hear the shape of the drum') is a classical problem in spectr al geometry with a broad range of implications and applications. While theoretic ally the answer to this question is negative (there exist examples of iso-spectr al but non-isometric manifolds), little is known about the practical possibility of using the spectrum for shape reconstruction and optimization. In this paper, we introduce a numerical procedure called isospectralization, consisting of def orming one shape to make its Laplacian spectrum match that of another. We implem ent the isospectralization procedure using modern differentiable programming tec hniques and exemplify its applications in some of the classical and notoriously hard problems in geometry processing, computer vision, and graphics such as shap e reconstruction, pose and style transfer, and dense deformable correspondence.

****

Speech2Face: Learning the Face Behind a Voice

Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Mi chael Rubinstein, Wojciech Matusik; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2019, pp. 7539-7548

How much can we infer about a person's looks from the way they speak? In this pa per, we study the task of reconstructing a facial image of a person from a short audio recording of that person speaking. We design and train a deep neural netw ork to perform this task using millions of natural Internet/Youtube videos of pe ople speaking. During training, our model learns voice-face correlations that al low it to produce images that capture various physical attributes of the speaker s such as age, gender and ethnicity. This is done in a self-supervised manner, b y utilizing the natural co-occurrence of faces and speech in Internet videos, wi thout the need to model attributes explicitly. We evaluate and numerically quant ify how--and in what manner--our Speech2Face reconstructions, obtained directly from audio, resemble the true face images of the speakers.

****

Joint Manifold Diffusion for Combining Predictions on Decoupled Observations

Kwang In Kim, Hyung Jin Chang; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2019, pp. 7549-7557

We present a new predictor combination algorithm that improves a given task pred ictor based on potentially relevant reference predictors. Existing approaches ar e limited in that, to discover the underlying task dependence, they either requi re known parametric forms of all predictors or access to a single fixed dataset on which all predictors are jointly evaluated. To overcome these limitations, we design a new non-parametric task dependence estimation procedure that automatic ally aligns evaluations of heterogeneous predictors across disjoint feature sets . Our algorithm is instantiated as a robust manifold diffusion process that join tly refines the estimated predictor alignments and the corresponding task depend ence. We apply this algorithm to the relative attributes ranking problem and dem onstrate that it not only broadens the application range of predictor combinatio n approaches but also outperforms existing methods even when applied to classica l predictor combination settings.

****

Audio Visual Scene-Aware Dialog

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, Devi Parikh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7558-7567

We introduce the task of scene-aware dialog. Our goal is to generate a complete and natural response to a question about a scene, given video and audio of the scene and the history of previous turns in the dialog. To answer successfully, agents must ground concepts from the question in the video while leveraging contextual cues from the dialog history. To benchmark this task, we introduce the Audio Visual Scene-Aware Dialog (AVSD) Dataset. For each of more than 11,000 videos of human actions from the Charades dataset, our dataset contains a dialog about the video, plus a final summary of the video by one of the dialog participants. We train several baseline systems for this task and evaluate the performance of the trained models using both qualitative and quantitative metrics. Our results indicate that models must utilize all the available inputs (video, audio, question, and dialog history) to perform best on this dataset.

****

Learning to Minify Photometric Stereo

Junxuan Li, Antonio Robles-Kelly, Shaodi You, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7568-7576

Photometric stereo estimates the surface normal given a set of images acquired under different illumination conditions. To deal with diverse factors involved in the image formation process, recent photometric stereo methods demand a large number of images as input. We propose a method that can dramatically decrease the demands on the number of images by learning the most informative ones under different illumination conditions. To this end, we use a deep learning framework to automatically learn the critical illumination conditions required at input. Furthermore, we present an occlusion layer that can synthesize cast shadows, which effectively improves the estimation accuracy. We assess our method on challenging real-world conditions, where we outperform techniques elsewhere in the literature with a significantly reduced number of light conditions.

****

Reflective and Fluorescent Separation Under Narrow-Band Illumination

Koji Koyamatsu, Daichi Hidaka, Takahiro Okabe, Hendrik P. A. Lensch; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7577-7585

In this paper, we address the separation of reflective and fluorescent components in RGB images taken under narrow-band light sources such as LEDs. First, we show that the fluorescent color per pixel can be estimated from at least two images under different light source colors, because the observed color at a surface point is represented by a convex combination of the light source color and the illumination-invariant fluorescent color. Second, we propose a method for robustly estimating the fluorescent color via MAP estimation by taking the prior knowledge with respect to fluorescent colors into consideration. We conducted a number of experiments by using both synthetic and real images, and confirmed that our proposed method works better than the closely related state-of-the-art method and enables us to separate reflective and fluorescent components even from a single image. Furthermore, we demonstrate that our method is effective for applications such as image-based material editing and relighting.

****

Depth From a Polarisation + RGB Stereo Pair

Dizhong Zhu, William A. P. Smith; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7586-7595

In this paper, we propose a hybrid depth imaging system in which a polarisation camera is augmented by a second image from a standard digital camera. For this modest increase in equipment complexity over conventional shape-from-polarisation, we obtain a number of benefits that enable us to overcome longstanding problems with the polarisation shape cue. The stereo cue provides a depth map which, although coarse, is metrically accurate. This is used as a guide surface for disam

biguation of the polarisation surface normal estimates using a higher order grap
hical model. In turn, these are used to estimate diffuse albedo. By extending a
previous shape-from-polarisation method to the perspective case, we show how to
compute dense, detailed maps of absolute depth, while retaining a linear formula
tion. We show that our hybrid method is able to recover dense 3D geometry that i
s superior to state-of-the-art shape-from-polarisation or two view stereo alone.
****

Rethinking the Evaluation of Video Summaries
Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkila; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 75
96-7604
Video summarization is a technique to create a short skim of the original video
while preserving the main stories/content. There exists a substantial interest i
n automatizing this process due to the rapid growth of the available material. T
he recent progress has been facilitated by public benchmark datasets, which enab
le easy and fair comparison of methods. Currently the established evaluation pro
tocol is to compare the generated summary with respect to a set of reference sum
maries provided by the dataset. In this paper, we will provide in-depth assessme
nt of this pipeline using two popular benchmark datasets. Surprisingly, we obser
ve that randomly generated summaries achieve comparable or better performance to
 the state-of-the-art. In some cases, the random summaries outperform even the h
uman generated summaries in leave-one-out experiments. Moreover, it turns out th
at the video segmentation, which is often considered as a fixed pre-processing m
ethod, has the most significant impact on the performance measure. Based on our
observations, we propose alternative approaches for assessing the importance sco
res as well as an intuitive visualization of correlation between the estimated s
coring and human annotations.
****

What Object Should I Use? - Task Driven Object Detection
Johann Sawatzky, Yaser Souri, Christian Grund, Jurgen Gall; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p
p. 7605-7614
When humans have to solve everyday tasks, they simply pick the objects that are
most suitable. While the question which object should one use for a specific tas
k sounds trivial for humans, it is very difficult to answer for robots or other
autonomous systems. This issue, however, is not addressed by current benchmarks
for object detection that focus on detecting object categories. We therefore int
roduce the COCO-Tasks dataset which comprises about 40,000 images where the most
 suitable objects for 14 tasks have been annotated. We furthermore propose an ap
proach that detects the most suitable objects for a given task. The approach bui
lds on a Gated Graph Neural Network to exploit the appearance of each object as
well as the global context of all present objects in the scene. In our experimen
ts, we show that the proposed approach outperforms other approaches that are eva
luated on the dataset like classification or ranking approaches.
****

Triangulation Learning Network: From Monocular to Stereo 3D Object Detection
Zengyi Qin, Jinglu Wang, Yan Lu; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2019, pp. 7615-7623
In this paper, we study the problem of 3D object detection from stereo images, i
n which the key challenge is how to effectively utilize stereo information. Diff
erent from previous methods using pixel-level depth maps, we propose to employ 3
D anchors to explicitly construct object-level correspondences between the regio
ns of interest in stereo images, from which the deep neural network learns to de
tect and triangulate the targeted object in 3D space. We also introduce a cost-e
fficient channel reweighting strategy that enhances representational features an
d weakens noisy signals to facilitate the learning process. All of these are fle
xibly integrated into a solid baseline detector that inputs monocular images. We
 demonstrate that both the monocular baseline and the stereo triangulation learn
ing network outperform the prior state-of-the-arts in 3D object detection and lo
calization on the challenging KITTI dataset.

****
Connecting the Dots: Learning Representations for Active Monocular Depth Estimation
Gernot Riegler, Yiyi Liao, Simon Donne, Vladlen Koltun, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7624-7633
We propose a technique for depth estimation with a monocular structured-light camera, i.e., a calibrated stereo set-up with one camera and one laser projector. Instead of formulating the depth estimation via a correspondence search problem, we show that a simple convolutional architecture is sufficient for high-quality disparity estimates in this setting. As accurate ground-truth is hard to obtain, we train our model in a self-supervised fashion with a combination of photometric and geometric losses. Further, we demonstrate that the projected pattern of the structured light sensor can be reliably separated from the ambient information. This can then be used to improve depth boundaries in a weakly supervised fashion by modeling the joint statistics of image and depth edges. The model trained in this fashion compares favorably to the state-of-the-art on challenging synthetic and real-world datasets. In addition, we contribute a novel simulator, which allows to benchmark active depth prediction algorithms in controlled conditions.
****
Learning Non-Volumetric Depth Fusion Using Successive Reprojections
Simon Donne, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7634-7643
Given a set of input views, multi-view stereopsis techniques estimate depth maps to represent the 3D reconstruction of the scene; these are fused into a single, consistent, reconstruction -- most often a point cloud. In this work we propose to learn an auto-regressive depth refinement directly from data. While deep learning has improved the accuracy and speed of depth estimation significantly, learned MVS techniques remain limited to the planesweeping paradigm. We refine a set of input depth maps by successively reprojecting information from neighbouring views to leverage multi-view constraints. Compared to learning-based volumetric fusion techniques, an image-based representation allows significantly more detailed reconstructions; compared to traditional point-based techniques, our method learns noise suppression and surface completion in a data-driven fashion. Due to the limited availability of high-quality reconstruction datasets with ground truth, we introduce two novel synthetic datasets to (pre-)train our network. Our approach is able to improve both the output depth maps and the reconstructed point cloud, for both learned and traditional depth estimation front-ends, on both synthetic and real data.
****
Stereo R-CNN Based 3D Object Detection for Autonomous Driving
Peiliang Li, Xiaozhi Chen, Shaojie Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7644-7652
We propose a 3D object detection method for autonomous driving by fully exploiting the sparse and dense, semantic and geometry information in stereo imagery. Our method, called Stereo R-CNN, extends Faster R-CNN for stereo inputs to simultaneously detect and associate object in left and right images. We add extra branches after stereo Region Proposal Network (RPN) to predict sparse keypoints, viewpoints, and object dimensions, which are combined with 2D left-right boxes to calculate a coarse 3D object bounding box. We then recover the accurate 3D bounding box by a region-based photometric alignment using left and right RoIs. Our method does not require depth input and 3D position supervision, however, outperforms all existing fully supervised image-based methods. Experiments on the challenging KITTI dataset show that our method outperforms the state-of-the-art stereo-based method by around 30% AP on both 3D detection and 3D localization tasks. Code will be made publicly available.
****
Hybrid Scene Compression for Visual Localization
Federico Camposeco, Andrea Cohen, Marc Pollefeys, Torsten Sattler; Proceeding

s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7653-7662
Localizing an image w.r.t. a 3D scene model represents a core task for many comp uter vision applications. An increasing number of real-world applications of vis ual localization on mobile devices, e.g., Augmented Reality or autonomous robots such as drones or self-driving cars, demand localization approaches to minimize storage and bandwidth requirements. Compressing the 3D models used for localiza tion thus becomes a practical necessity. In this work, we introduce a new hybrid compression algorithm that uses a given memory limit in a more effective way. R ather than treating all 3D points equally, it represents a small set of points w ith full appearance information and an additional, larger set of points with com pressed information. This enables our approach to obtain a more complete scene r epresentation without increasing the memory requirements, leading to a superior performance compared to previous compression schemes. As part of our contributio n, we show how to handle ambiguous matches arising from point compression during RANSAC. Besides outperforming previous compression techniques in terms of pose accuracy under the same memory constraints, our compression scheme itself is al so more efficient. Furthermore, the localization rates and accuracy obtained wit h our approach are comparable to state-of-the-art feature-based methods, while u sing a small fraction of the memory.
****

MMFace: A Multi-Metric Regression Network for Unconstrained Face Reconstruction
Hongwei Yi, Chen Li, Qiong Cao, Xiaoyong Shen, Sheng Li, Guoping Wang, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7663-7672
We propose to address the face reconstruction in the wild by using a multi-metri c regression network, MMFace, to align a 3D face morphable model (3DMM) to an in put image. The key idea is to utilize a volumetric sub-network to estimate an in termediate geometry representation, and a parametric sub-network to regress the 3DMM parameters. Our parametric sub-network consists of identity loss, expressio n loss, and pose loss which greatly improves the aligned geometry details by inc orporating high level loss functions directly defined in the 3DMM parametric spa ces. Our high-quality reconstruction is robust under large variations of express ions, poses, illumination conditions, and even with large partial occlusions. We evaluate our method by comparing the performance with state-of-the-art approach es on latest 3D face dataset LS3D-W and Florence. We achieve significant improve ments both quantitatively and qualitatively. Due to our high-quality reconstruct ion, our method can be easily extended to generate high-quality geometry sequenc es for video inputs.
****

3D Motion Decomposition for RGBD Future Dynamic Scene Synthesis
Xiaojuan Qi, Zhengzhe Liu, Qifeng Chen, Jiaya Jia; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7673-7 682
A future video is the 2D projection of a 3D scene with predicted camera and obje ct motion. Accurate future video prediction inherently requires understanding of 3D motion and geometry of a scene. In this paper, we propose a RGBD scene forec asting model with 3D motion decomposition. We predict ego-motion and foreground motion that are combined to generate a future 3D dynamic scene, which is then pr ojected into a 2D image plane to synthesize future motion, RGB images and depth maps. Optional semantic maps can be integrated. Experimental results on KITTI an d Driving datasets show that our model outperforms other state-of-the- arts in f orecasting future RGBD dynamic scenes.
****

Single Image Depth Estimation Trained via Depth From Defocus Cues
Shir Gur, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7683-7692
Estimating depth from a single RGB images is a fundamental task in computer visi on, which is most directly solved using supervised deep learning. In the field o f unsupervised learning of depth from a single RGB image, depth is not given exp

licitly. Existing work in the field receives either a stereo pair, a monocular v ideo, or multiple views, and, using losses that are based on structure-from-moti on, trains a depth estimation network. In this work, we rely, instead of differe nt views, on depth from focus cues. Learning is based on a novel Point Spread Fu nction convolutional layer, which applies location specific kernels that arise f rom the Circle-Of-Confusion in each image location. We evaluate our method on da ta derived from five common datasets for depth estimation and lightfield images, and present results that are on par with supervised methods on KITTI and Make3D datasets and outperform unsupervised learning approaches. Since the phenomenon of depth from defocus is not dataset specific, we hypothesize that learning base d on it would overfit less to the specific content in each dataset. Our experime nts show that this is indeed the case, and an estimator learned on one dataset u sing our method provides better results on other datasets, than the directly sup ervised methods.

****

RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Comp letion

Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, Ian Reid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2019, pp. 7693-7702

RGB images differentiate from depth as they carry more details about the color a nd texture information, which can be utilized as a vital complement to depth for boosting the performance of 3D semantic scene completion (SSC). SSC is composed of 3D shape completion (SC) and semantic scene labeling while most of the exist ing approaches use depth as the sole input which causes the performance bottlene ck. Moreover, the state-of-the-art methods employ 3D CNNs which have cumbersome networks and tremendous parameters. We introduce a light-weight Dimensional Deco mposition Residual network (DDR) for 3D dense prediction tasks. The novel factor ized convolution layer is effective for reducing the network parameters, and the proposed multi-scale fusion mechanism for depth and color image can improve the completion and segmentation accuracy simultaneously. Our method demonstrates ex cellent performance on two public datasets. Compared with the latest method SSCN et, we achieve 5.9% gains in SC-IoU and 5.7% gains in SSC-IOU, albeit with only 21% network parameters and 16.6% FLOPs employed compared with that of SSCNet.

****

Neural Scene Decomposition for Multi-Person Motion Capture

Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, Pasca l Fua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2019, pp. 7703-7713

Learning general image representations has proven key to the success of many com puter vision tasks. For example, many approaches to image understanding problems rely on deep networks that were initially trained on ImageNet, mostly because t he learned features are a valuable starting point to learn from limited labeled data. However, when it comes to 3D motion capture of multiple people, these feat ures are only of limited use. In this paper, we therefore propose an approach to learning features that are useful for this purpose. To this end, we introduce a self-supervised approach to learning what we call a neural scene decompositio n (NSD) that can be exploited for 3D pose estimation. NSD comprises three layer s of abstraction to represent human subjects: spatial layout in terms of boundin g-boxes and relative depth; a 2D shape representation in terms of an instance se gmentation mask; and subject-specific appearance and 3D pose information. By exp loiting self-supervision coming from multiview data, our NSD model can be traine d end-to-end without any 2D or 3D supervision. In contrast to previous approache s, it works for multiple persons and full-frame images. Because it encodes 3D ge ometry, NSD can then be effectively leveraged to train a 3D pose estimation netw ork from small amounts of annotated data.

****

Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition

Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, Jun Zh u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni

tion (CVPR), 2019, pp. 7714-7722
Face recognition has obtained remarkable progress in recent years due to the great improvement of deep convolutional neural networks (CNNs). However, deep CNNs are vulnerable to adversarial examples, which can cause fateful consequences in real-world face recognition applications with security-sensitive purposes. Adversarial attacks are widely studied as they can identify the vulnerability of the models before they are deployed. In this paper, we evaluate the robustness of state-of-the-art face recognition models in the decision-based black-box attack setting, where the attackers have no access to the model parameters and gradients, but can only acquire hard-label predictions by sending queries to the target model. This attack setting is more practical in real-world face recognition systems. To improve the efficiency of previous methods, we propose an evolutionary attack algorithm, which can model the local geometry of the search directions and reduce the dimension of the search space. Extensive experiments demonstrate the effectiveness of the proposed method that induces a minimum perturbation to an input face image with fewer queries. We also apply the proposed method to attack a real-world face recognition system successfully.
****

FA-RPN: Floating Region Proposals for Face Detection
Mahyar Najibi, Bharat Singh, Larry S. Davis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7723-7732
We propose a novel approach for generating region proposals for performing face detection. Instead of classifying anchor boxes using features from a pixel in the convolutional feature map, we adopt a pooling-based approach for generating region proposals. However, pooling hundreds of thousands of anchors which are evaluated for generating proposals becomes a computational bottleneck during inference. To this end, an efficient anchor placement strategy for reducing the number of anchor-boxes is proposed. We then show that proposals generated by our network (Floating Anchor Region Proposal Network, FA-RPN) are better than RPN for generating region proposals for face detection. We discuss several beneficial features of FA-RPN proposals (which can be enabled without re-training) like iterative refinement, placement of fractional anchors and changing size/shape of anchors. Our face detector based on FA-RPN obtains 89.4% mAP with a ResNet-50 backbone on the WIDER dataset.
****

Bayesian Hierarchical Dynamic Model for Human Action Recognition
Rui Zhao, Wanru Xu, Hui Su, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7733-7742
Human action recognition remains as a challenging task partially due to the presence of large variations in the execution of action. To address this issue, we propose a probabilistic model called Hierarchical Dynamic Model (HDM). Leveraging on Bayesian framework, the model parameters are allowed to vary across different sequences of data, which increase the capacity of the model to adapt to intra-class variations on both spatial and temporal extent of actions. Meanwhile, the generative learning process allows the model to preserve the distinctive dynamic pattern for each action class. Through Bayesian inference, we are able to quantify the uncertainty of the classification, providing insight during the decision process. Compared to state-of-the-art methods, our method not only achieves competitive recognition performance within individual dataset but also shows better generalization capability across different datasets. Experiments conducted on data with missing values also show the robustness of the proposed method.
****

Mixed Effects Neural Networks (MeNets) With Applications to Gaze Estimation
Yunyang Xiong, Hyunwoo J. Kim, Vikas Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7743-7752
There is much interest in computer vision to utilize commodity hardware for gaze estimation. A number of papers have shown that algorithms based on deep convolutional architectures are approaching accuracies where streaming data from mass-market devices can offer good gaze tracking performance, although a gap still remains between what is possible and the performance users will expect in real depl

oyments. We observe that one obvious avenue for improvement relates to a gap between some basic technical assumptions behind most existing approaches and the statistical properties of the data used for training. Specifically, most training datasets involve tens of users with a few hundreds (or more) repeated acquisitions per user. The non i.i.d. nature of this data suggests better estimation may be possible if the model explicitly made use of such "repeated measurements" from each user as is commonly done in classical statistical analysis using so-called mixed effects models. The goal of this paper is to adapt these "mixed effects" ideas from statistics within a deep neural network architecture for gaze estimation, based on eye images. Such a formulation seeks to specifically utilize information regarding the hierarchical structure of the training data -- each node in the hierarchy is a user who provides tens or hundreds of repeated samples. This modification yields an architecture that offers state of the art performance on various publicly available datasets improving results by 10-20%.
****

3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training

Dario Pavllo, Christoph Feichtenhofer, David Grangier, Michael Auli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7753-7762

In this work, we demonstrate that 3D poses in video can be effectively estimated with a fully convolutional model based on dilated temporal convolutions over 2D keypoints. We also introduce back-projection, a simple and effective semi-supervised training method that leverages unlabeled video data. We start with predicted 2D keypoints for unlabeled video, then estimate 3D poses and finally back-project to the input 2D keypoints. In the supervised setting, our fully-convolutional model outperforms the previous best result from the literature by 6 mm mean per-joint position error on Human3.6M, corresponding to an error reduction of 11%, and the model also shows significant improvements on HumanEva-I. Moreover, experiments with back-projection show that it comfortably outperforms previous state-of-the-art results in semi-supervised settings where labeled data is scarce. Code and models are available at https://github.com/facebookresearch/VideoPose3D
****

Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7763-7772

The estimation of 3D face shape from a single image must be robust to variations in lighting, head pose, expression, facial hair, makeup, and occlusions. Robustness requires a large training set of in-the-wild images, which by construction, lack ground truth 3D shape. To train a network without any 2D-to-3D supervision, we present RingNet, which learns to compute 3D face shape from a single image. Our key observation is that an individual's face shape is constant across images, regardless of expression, pose, lighting, etc. RingNet leverages multiple images of a person and automatically detected 2D face features. It uses a novel loss that encourages the face shape to be similar when the identity is the same and different for different people. We achieve invariance to expression by representing the face using the FLAME model. Once trained, our method takes a single image and outputs the parameters of FLAME, which can be readily animated. Additionally we create a new database of faces "not quite in-the-wild" (NoW) with 3D head scans and high-resolution images of the subjects in a wide variety of conditions. We evaluate publicly available methods and find that RingNet is more accurate than methods that use 3D supervision. The dataset, model, and results are available for research purposes at http://ringnet.is.tuebingen.mpg.de.
****

PoseFix: Model-Agnostic General Human Pose Refinement Network

Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7773-7781

Multi-person pose estimation from a 2D image is an essential technique for human

behavior understanding. In this paper, we propose a human pose refinement netwo rk that estimates a refined pose from a tuple of an input image and input pose. The pose refinement was performed mainly through an end-to-end trainable multi-s tage architecture in previous methods. However, they are highly dependent on pos e estimation models and require careful model design. By contrast, we propose a model-agnostic pose refinement method. According to a recent study, state-of-the -art 2D human pose estimation methods have similar error distributions. We use t his error statistics as prior information to generate synthetic poses and use th e synthesized poses to train our model. In the testing stage, pose estimation re sults of any other methods can be input to the proposed method. Moreover, the pr oposed model does not require code or knowledge about other methods, which allow s it to be easily used in the post-processing step. We show that the proposed ap proach achieves better performance than the conventional multi-stage refinement models and consistently improves the performance of various state-of-the-art pos e estimation methods on the commonly used benchmark. The code is available in (h ttps://github.com/mks0601/PoseFix_RELEASE).

****

RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation

Bastian Wandt, Bodo Rosenhahn; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2019, pp. 7782-7791

This paper addresses the problem of 3D human pose estimation from single images. While for a long time human skeletons were parameterized and fitted to the obse rvation by satisfying a reprojection error, nowadays researchers directly use ne ural networks to infer the 3D pose from the observations. However, most of these approaches ignore the fact that a reprojection constraint has to be satisfied a nd are sensitive to overfitting. We tackle the overfitting problem by ignoring 2 D to 3D correspondences. This efficiently avoids a simple memorization of the tr aining data and allows for a weakly supervised training. One part of the propose d reprojection network (RepNet) learns a mapping from a distribution of 2D poses to a distribution of 3D poses using an adversarial training approach. Another p art of the network estimates the camera. This allows for the definition of a net work layer that performs the reprojection of the estimated 3D pose back to 2D wh ich results in a reprojection loss function. Our experiments show that RepNet generalizes well to unknown data and outperforms state-of-the-art methods when a pplied to unseen data. Moreover, our implementation runs in real-time on a stand ard desktop PC.

****

Fast and Robust Multi-Person 3D Pose Estimation From Multiple Views

Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 19, pp. 7792-7801

This paper addresses the problem of 3D pose estimation for multiple people in a few calibrated camera views. The main challenge of this problem is to find the c ross-view correspondences among noisy and incomplete 2D pose predictions. Most p revious methods address this challenge by directly reasoning in 3D using a picto rial structure model, which is inefficient due to the huge state space. We propo se a fast and robust approach to solve this problem. Our key idea is to use a mu lti-way matching algorithm to cluster the detected 2D poses in all views. Each r esulting cluster encodes 2D poses of the same person across different views and consistent correspondences across the keypoints, from which the 3D pose of each person can be effectively inferred. The proposed convex optimization based multi -way matching algorithm is efficient and robust against missing and false detect ions, without knowing the number of people in the scene. Moreover, we propose to combine geometric and appearance cues for cross-view matching. The proposed app roach achieves significant performance gains from the state-of-the-art (96.3% vs . 90.6% and 96.9% vs. 88% on the Campus and Shelf datasets, respectively), while being efficient for real-time applications.

****

Face-Focused Cross-Stream Network for Deception Detection in Videos

Mingyu Ding, An Zhao, Zhiwu Lu, Tao Xiang, Ji-Rong Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7802-7811

Automated deception detection (ADD) from real-life videos is a challenging task. It specifically needs to address two problems: (1) Both face and body contain useful cues regarding whether a subject is deceptive. How to effectively fuse the two is thus key to the effectiveness of an ADD model. (2) Real-life deceptive samples are hard to collect; learning with limited training data thus challenges most deep learning based ADD models. In this work, both problems are addressed. Specifically, for face-body multimodal learning, a novel face-focused cross-stream network (FFCSN) is proposed. It differs significantly from the popular two-stream networks in that: (a) face detection is added into the spatial stream to capture the facial expressions explicitly, and (b) correlation learning is performed across the spatial and temporal streams for joint deep feature learning across both face and body. To address the training data scarcity problem, our FFCSN model is trained with both meta learning and adversarial learning. Extensive experiments show that our FFCSN model achieves state-of-the-art results. Further, the proposed FFCSN model as well as its robust training strategy are shown to be generally applicable to other human-centric video analysis tasks such as emotion recognition from user-generated videos.

****

Unequal-Training for Deep Face Recognition With Long-Tailed Noisy Data

Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, Yaohai Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7812-7821

Large-scale face datasets usually exhibit a massive number of classes, a long-tailed distribution, and severe label noise, which undoubtedly aggravate the difficulty of training. In this paper, we propose a training strategy that treats the head data and the tail data in an unequal way, accompanying with noise-robust loss functions, to take full advantage of their respective characteristics. Specifically, the unequal-training framework provides two training data streams: the first stream applies the head data to learn discriminative face representation supervised by Noise Resistance loss; the second stream applies the tail data to learn auxiliary information by gradually mining the stable discriminative information from confusing tail classes. Consequently, both training streams offer complementary information to deep feature learning. Extensive experiments have demonstrated the effectiveness of the new unequal-training framework and loss functions. Better yet, our method could save a significant amount of GPU memory. With our method, we achieve the best result on MegaFace Challenge 2 (MF2) given a large-scale noisy training data set.

****

T-Net: Parametrizing Fully Convolutional Nets With a Single High-Order Tensor

Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, Maja Pantic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7822-7831

Recent findings indicate that over-parametrization, while crucial for successfully training deep neural networks, also introduces large amounts of redundancy. Tensor methods have the potential to efficiently parametrize over-complete representations by leveraging this redundancy. In this paper, we propose to fully parametrize Convolutional Neural Networks (CNNs) with a single high-order, low-rank tensor. Previous works on network tensorization have focused on parametrizing individual layers (convolutional or fully connected) only, and perform the tensorization layer-by-layer separately. In contrast, we propose to jointly capture the full structure of a neural network by parametrizing it with a single high-order tensor, the modes of which represent each of the architectural design parameters of the network (e.g. number of convolutional blocks, depth, number of stacks, input features, etc). This parametrization allows to regularize the whole network and drastically reduce the number of parameters. Our model is end-to-end trainable and the low-rank structure imposed on the weight tensor acts as an implicit regularization. We study the case of networks with rich structure, namely Fully

Convolutional Networks (FCNs), which we propose to parametrize with a single 8th-order tensor. We show that our approach can achieve superior performance with small compression rates, and attain high compression rates with negligible drop in accuracy for the challenging task of human pose estimation.
****

Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss
Lele Chen, Ross K. Maddox, Zhiyao Duan, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7832-7841
We devise a cascade GAN approach to generate talking face video, which is robust to different face shapes, view angles, facial characteristics, and noisy audio conditions. Instead of learning a direct mapping from audio to video frames, we propose first to transfer audio to high-level structure, i.e., the facial landmarks, and then to generate video frames conditioned on the landmarks. Compared to a direct audio-to-image approach, our cascade approach avoids fitting spurious correlations between audiovisual signals that are irrelevant to the speech content. We, humans, are sensitive to temporal discontinuities and subtle artifacts in video. To avoid those pixel jittering problems and to enforce the network to focus on audiovisual-correlated regions, we propose a novel dynamically adjustable pixel-wise loss with an attention mechanism. Furthermore, to generate a sharper image with well-synchronized facial movements, we propose a novel regression-based discriminator structure, which considers sequence-level information along with frame-level information. Thoughtful experiments on several datasets and real-world samples demonstrate significantly better results obtained by our method than the state-of-the-art methods in both quantitative and qualitative comparisons.
****

Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video
Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7842-7851
Abnormal event detection in video is a challenging vision problem. Most existing approaches formulate abnormal event detection as an outlier detection task, due to the scarcity of anomalous data during training. Because of the lack of prior information regarding abnormal events, these methods are not fully-equipped to differentiate between normal and abnormal events. In this work, we formalize abnormal event detection as a one-versus-rest binary classification problem. Our contribution is two-fold. First, we introduce an unsupervised feature learning framework based on object-centric convolutional auto-encoders to encode both motion and appearance information. Second, we propose a supervised classification approach based on clustering the training samples into normality clusters. A one-versus-rest abnormal event classifier is then employed to separate each normality cluster from the rest. For the purpose of training the classifier, the other clusters act as dummy anomalies. During inference, an object is labeled as abnormal if the highest classification score assigned by the one-versus-rest classifiers is negative. Comprehensive experiments are performed on four benchmarks: Avenue, ShanghaiTech, UCSD and UMN. Our approach provides superior results on all four data sets. On the large-scale ShanghaiTech data set, our method provides an absolute gain of 8.4% in terms of frame-level AUC compared to the state-of-the-art method.
****

DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition
Toby Perrett, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7852-7861
Domain alignment in convolutional networks aims to learn the degree of layer-specific feature alignment beneficial to the joint learning of source and target datasets. While increasingly popular in convolutional networks, there have been no previous attempts to achieve domain alignment in recurrent networks. Similar to spatial features, both source and target domains are likely to exhibit temporal

dependencies that can be jointly learnt and aligned. In this paper we introduce Dual-Domain LSTM (DDLSTM), an architecture that is able to learn temporal dependencies from two domains concurrently. It performs cross-contaminated batch normalisation on both input-to-hidden and hidden-to-hidden weights, and learns the parameters for cross-contamination, for both single-layer and multi-layer LSTM architectures. We evaluate DDLSTM on frame-level action recognition using three datasets, taking a pair at a time, and report an average increase in accuracy of 3.5%. The proposed DDLSTM architecture outperforms standard, fine-tuned, and batch-normalised LSTMs.

****

The Pros and Cons: Rank-Aware Temporal Attention for Skill Determination in Long Videos

Hazel Doughty, Walterio Mayol-Cuevas, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7862-7871

We present a new model to determine relative skill from long videos, through learnable temporal attention modules. Skill determination is formulated as a ranking problem, making it suitable for common and generic tasks. However, for long videos, parts of the video are irrelevant for assessing skill, and there may be variability in the skill exhibited throughout a video. We therefore propose a method which assesses the relative overall level of skill in a long video by attending to its skill-relevant parts. Our approach trains temporal attention modules, learned with only video-level supervision, using a novel rank-aware loss function. In addition to attending to task-relevant video parts, our proposed loss jointly trains two attention modules to separately attend to video parts which are indicative of higher (pros) and lower (cons) skill. We evaluate our approach on the EPIC-Skills dataset and additionally annotate a larger dataset from YouTube videos for skill determination with five previously unexplored tasks. Our method outperforms previous approaches and classic softmax attention on both datasets by over 4% pairwise accuracy, and as much as 12% on individual tasks. We also demonstrate our model's ability to attend to rank-aware parts of the video.

****

Collaborative Spatiotemporal Feature Learning for Video Action Recognition

Chao Li, Qiaoyong Zhong, Di Xie, Shiliang Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7872-7881

Spatiotemporal feature learning is of central importance for action recognition in videos. Existing deep neural network models either learn spatial and temporal features independently (C2D) or jointly with unconstrained parameters (C3D). In this paper, we propose a novel neural operation which encodes spatiotemporal features collaboratively by imposing a weight-sharing constraint on the learnable parameters. In particular, we perform 2D convolution along three orthogonal views of volumetric video data, which learns spatial appearance and temporal motion cues respectively. By sharing the convolution kernels of different views, spatial and temporal features are collaboratively learned and thus benefit from each other. The complementary features are subsequently fused by a weighted summation whose coefficients are learned end-to-end. Our approach achieves state-of-the-art performance on large-scale benchmarks and won the 1st place in the Moments in Time Challenge 2018. Moreover, based on the learned coefficients of different views, we are able to quantify the contributions of spatial and temporal features. This analysis sheds light on interpretability of the model and may also guide the future design of algorithm for video recognition.

****

MARS: Motion-Augmented RGB Stream for Action Recognition

Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7882-7891

Most state-of-the-art methods for action recognition consist of a two-stream architecture with 3D convolutions: an appearance stream for RGB frames and a motion stream for optical flow frames. Although combining flow with RGB improves the performance, the cost of computing accurate optical flow is high, and increases

action recognition latency. This limits the usage of two-stream approaches in re
al-world applications requiring low latency. In this paper, we introduce two lea
rning approaches to train a standard 3D CNN, operating on RGB frames, that mimic
s the motion stream, and as a result avoids flow computation at test time. Firs
t, by minimizing a feature-based loss compared to the Flow stream, we show that
the network reproduces the motion stream with high fidelity. Second, to leverag
e both appearance and motion information effectively, we train with a linear com
bination of the feature-based loss and the standard cross-entropy loss for actio
n recognition. We denote the stream trained using this combined loss as Motion-A
ugmented RGB Stream (MARS). As a single stream, MARS performs better than RGB or
Flow alone, for instance with 72.7% accuracy on Kinetics compared to 72.0% and
65.6% with RGB and Flow streams respectively.
****

Convolutional Relational Machine for Group Activity Recognition
Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, Alexandre Alahi;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2019, pp. 7892-7901
We present an end-to-end deep Convolutional Neural Network called Convolutional
Relational Machine (CRM) for recognizing group activities that utilizes the info
rmation in spatial relations between individual persons in image or video. It le
arns to produce an intermediate spatial representation (activity map) based on i
ndividual and group activities. A multi-stage refinement component is responsibl
e for decreasing the incorrect predictions in the activity map. Finally, an a
ggregation component uses the refined information to recognize group activities.
Experimental results demonstrate the constructive contribution of the informati
on extracted and represented in the form of the activity map. CRM shows advantag
es over state-of-the-art models on Volleyball and Collective Activity datasets.
****

Video Summarization by Learning From Unpaired Data
Mrigank Rochan, Yang Wang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2019, pp. 7902-7911
We consider the problem of video summarization. Given an input raw video, the go
al is to select a small subset of key frames from the input video to create a sh
orter summary video that best describes the content of the original video. Most
of the current state-of-the-art video summarization approaches use supervised le
arning and require labeled training data. Each training instance consists of a r
aw input video and its ground truth summary video curated by human annotators. H
owever, it is very expensive and difficult to create such labeled training examp
les. To address this limitation, we propose a novel formulation to learn video s
ummarization from unpaired data. We present an approach that learns to generate
optimal video summaries using a set of raw videos (V) and a set of summary video
s (S), where there exists no correspondence between V and S. We argue that this
type of data is much easier to collect. Our model aims to learn a mapping functi
on F : V -> S such that the distribution of resultant summary videos from F(V) i
s similar to the distribution of S with the help of an adversarial objective. In
addition, we enforce a diversity constraint on F(V) to ensure that the generate
d video summaries are visually diverse. Experimental results on two benchmark da
tasets indicate that our proposed approach significantly outperforms other alter
native methods.
****

Skeleton-Based Action Recognition With Directed Graph Neural Networks
Lei Shi, Yifan Zhang, Jian Cheng, Hanqing Lu; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7912-7921
The skeleton data have been widely used for the action recognition tasks since t
hey can robustly accommodate dynamic circumstances and complex backgrounds. In e
xisting methods, both the joint and bone information in skeleton data have been
proved to be of great help for action recognition tasks. However, how to incorpo
rate these two types of data to best take advantage of the relationship between
joints and bones remains a problem to be solved. In this work, we represent the
skeleton data as a directed acyclic graph based on the kinematic dependency betw

een the joints and bones in the natural human body. A novel directed graph neura
l network is designed specially to extract the information of joints, bones and
their relations and make prediction based on the extracted features. In addition
, to better fit the action recognition task, the topological structure of the gr
aph is made adaptive based on the training process, which brings notable improve
ment. Moreover, the motion information of the skeleton sequence is exploited and
 combined with the spatial information to further enhance the performance in a t
wo-stream framework. Our final model is tested on two large-scale datasets, NTU-
RGBD and Skeleton-Kinetics, and exceeds state-of-the-art performance on both of
them.
****

PA3D: Pose-Action 3D Machine for Video Recognition
An Yan, Yali Wang, Zhifeng Li, Yu Qiao; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7922-7931
Recent studies have witnessed the successes of using 3D CNNs for video action re
cognition. However, most 3D models are built upon RGB and optical flow streams,
which may not fully exploit pose dynamics, i.e., an important cue of modeling hu
man actions. To fill this gap, we propose a concise Pose-Action 3D Machine (PA3D
), which can effectively encode multiple pose modalities within a unified 3D fra
mework, and consequently learn spatio-temporal pose representations for action r
ecognition. More specifically, we introduce a novel temporal pose convolution t
o aggregate spatial poses over frames. Unlike the classical temporal convolution
, our operation can explicitly learn the pose motions that are discriminative to
 recognize human actions. Extensive experiments on three popular benchmarks (i.e
., JHMDB, HMDB, and Charades) show that, PA3D outperforms the recent pose-based
approaches. Furthermore, PA3D is highly complementary to the recent 3D CNNs, e.g
., I3D. Multi-stream fusion achieves the state-of-the-art performance on all eva
luated data sets.
****

Deep Dual Relation Modeling for Egocentric Interaction Recognition
Haoxin Li, Yijun Cai, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7932-7941
Egocentric interaction recognition aims to recognize the camera wearer's interac
tions with the interactor who faces the camera wearer in egocentric videos. In s
uch a human-human interaction analysis problem, it is crucial to explore the rel
ations between the camera wearer and the interactor. However, most existing work
s directly model the interactions as a whole and lack modeling the relations bet
ween the two interacting persons. To exploit the strong relations for egocentric
 interaction recognition, we introduce a dual relation modeling framework which
learns to model the relations between the camera wearer and the interactor based
 on the individual action representations of the two persons. Specifically, we d
evelop a novel interactive LSTM module, the key component of our framework, to e
xplicitly model the relations between the two interacting persons based on their
 individual action representations, which are collaboratively learned with an in
teractor attention module and a global-local motion module. Experimental results
 on three egocentric interaction datasets show the effectiveness of our method a
nd advantage over state-of-the-arts.
****

MOTS: Multi-Object Tracking and Segmentation
Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balac
handar Gnana Sekar, Andreas Geiger, Bastian Leibe; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7942-79
51
This paper extends the popular task of multi-object tracking to multi-object tra
cking and segmentation (MOTS). Towards this goal, we create dense pixel-level an
notations for two existing tracking datasets using a semi-automatic annotation p
rocedure. Our new annotations comprise 65,213 pixel masks for 977 distinct objec
ts (cars and pedestrians) in 10,870 video frames. For evaluation, we extend exis
ting multi-object tracking metrics to this new task. Moreover, we propose a new
baseline method which jointly addresses detection, tracking, and segmentation wi

th a single convolutional network. We demonstrate the value of our datasets by a chieving improvements in performance when training on MOTS annotations. We belie ve that our datasets, metrics and baseline will become a valuable resource towar ds developing multi-object tracking approaches that go beyond 2D bounding boxes. We make our annotations, code, and models available at https://www.vision.rwth-aachen.de/page/mots.
****

Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking
Heng Fan, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2019, pp. 7952-7961
Recently, the region proposal networks (RPN) have been combined with the Siamese network for tracking, and shown excellent accuracy with high efficiency. Nevert heless, previously proposed one-stage Siamese-RPN trackers degenerate in presenc e of similar distractors and large scale variation. Addressing these issues, we propose a multi-stage tracking framework, Siamese Cascaded RPN (C-RPN), which co nsists of a sequence of RPNs cascaded from deep high-level to shallow low-level layers in a Siamese network. Compared to previous solutions, C-RPN has several a dvantages: (1) Each RPN is trained using the outputs of RPN in the previous stag e. Such process stimulates hard negative sampling, resulting in more balanced tr aining samples. Consequently, the RPNs are sequentially more discriminative in d istinguishing difficult background (i.e.,, similar distractors). (2) Multi-level features are fully leveraged through a novel feature transfer block (FTB) for e ach RPN, further improving the discriminability of C-RPN using both high-level s emantic and low-level spatial information. (3) With multiple steps of regression s, C-RPN progressively refines the location and shape of the target in each RPN with adjusted anchor boxes in the previous stage, which makes localization more accurate. C-RPN is trained end-to-end with the multi-task loss function. In infe rence, C-RPN is deployed as it is, without any temporal adaption, for real-time tracking. In extensive experiments on OTB-2013, OTB-2015, VOT-2016, VOT-2017, La SOT and TrackingNet, C-RPN consistently achieves state-of-the-art results and ru ns in real-time.
****

PointFlowNet: Learning Representations for Rigid Motion Estimation From Point Cl ouds
Aseem Behl, Despoina Paschalidou, Simon Donne, Andreas Geiger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 , pp. 7962-7971
Despite significant progress in image-based 3D scene flow estimation, the perfor mance of such approaches has not yet reached the fidelity required by many appli cations. Simultaneously, these applications are often not restricted to image-ba sed estimation: laser scanners provide a popular alternative to traditional came ras, for example in the context of self-driving cars, as they directly yield a 3 D point cloud. In this paper, we propose to estimate 3D motion from such unstruc tured point clouds using a deep neural network. In a single forward pass, our mo del jointly predicts 3D scene flow as well as the 3D bounding box and rigid body motion of objects in the scene. While the prospect of estimating 3D scene flow from unstructured point clouds is promising, it is also a challenging task. We show that the traditional global representation of rigid body motion prohibits i nference by CNNs, and propose a translation equivariant representation to circum vent this problem. For training our deep network, a large dataset is required. B ecause of this, we augment real scans from KITTI with virtual objects, realistic ally modeling occlusions and simulating sensor noise. A thorough comparison with classic and learning-based techniques highlights the robustness of the proposed approach.
****

Listen to the Image
Di Hu, Dong Wang, Xuelong Li, Feiping Nie, Qi Wang; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7972 -7981
Visual-to-auditory sensory substitution devices can assist the blind in sensing

the visual environment by translating the visual information into a sound patter
n. To improve the translation quality, the task performances of the blind are us
ually employed to evaluate different encoding schemes. In contrast to the toilso
me human-based assessment, we argue that machine model can be also developed for
 evaluation, and more efficient. To this end, we firstly propose two distinct cr
oss-modal perception model w.r.t. the late-blind and congenitally-blind cases, w
hich aim to generate concrete visual contents based on the translated sound. To
validate the functionality of proposed models, two novel optimization strategies
 w.r.t. the primary encoding scheme are presented. Further, we conduct sets of h
uman-based experiments to evaluate and compare them with the conducted machine-b
ased assessments in the cross-modal generation task. Their highly consistent res
ults w.r.t. different encoding schemes indicate that using machine model to acce
lerate optimization evaluation and reduce experimental cost is feasible to some
extent, which could dramatically promote the upgrading of encoding scheme then h
elp the blind to improve their visual perception ability.
****

Image Super-Resolution by Neural Texture Transfer
Zhifei Zhang, Zhaowen Wang, Zhe Lin, Hairong Qi; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7982-799
1
Due to the significant information loss in low-resolution (LR) images, it has be
come extremely challenging to further advance the state-of-the-art of single ima
ge super-resolution (SISR). Reference-based super-resolution (RefSR), on the oth
er hand, has proven to be promising in recovering high-resolution (HR) details w
hen a reference (Ref) image with similar content as that of the LR input is give
n. However, the quality of RefSR can degrade severely when Ref is less similar.
This paper aims to unleash the potential of RefSR by leveraging more texture det
ails from Ref images with stronger robustness even when irrelevant Ref images ar
e provided. Inspired by the recent work on image stylization, we formulate the R
efSR problem as neural texture transfer. We design an end-to-end deep model whic
h enriches HR details by adaptively transferring the texture from Ref images acc
ording to their textural similarity. Instead of matching content in the raw pixe
l space as done by previous methods, our key contribution is a multi-level match
ing conducted in the neural space. This matching scheme facilitates multi-scale
neural transfer that allows the model to benefit more from those semantically re
lated Ref patches, and gracefully degrade to SISR performance on the least relev
ant Ref inputs. We build a benchmark dataset for the general research of RefSR,
which contains Ref images paired with LR inputs with varying levels of similarit
y. Both quantitative and qualitative evaluations demonstrate the superiority of
our method over state-of-the-art.
****

Conditional Adversarial Generative Flow for Controllable Image Synthesis
Rui Liu, Yu Liu, Xinyu Gong, Xiaogang Wang, Hongsheng Li; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp
. 7992-8001
Flow-based generative models show great potential in image synthesis due to its
reversible pipeline and exact log-likelihood target, yet it suffers from weak ab
ility for conditional image synthesis, especially for multi-label or unaware con
ditions. This is because the potential distribution of image conditions is hard
to measure precisely from its latent variable z. In this paper, based on modelin
g a joint probabilistic density of an image and its conditions, we propose a nov
el flow-based generative model named conditional adversarial generative flow (CA
Glow). Instead of disentangling attributes from latent space, we blaze a new tra
il for learning an encoder to estimate the mapping from condition space to laten
t space in an adversarial manner. Given a specific condition c, CAGlow can encod
e it to a sampled z, and then enable robust conditional image synthesis in compl
ex situations like combining person identity with multiple attributes. The propo
sed CAGlow can be implemented in both supervised and unsupervised manners, thus
can synthesize images with conditional information like categories, attributes,
and even some unknown properties. Extensive experiments show that CAGlow ensures

the independence of different conditions and outperforms regular Glow to a significant extent.
****

How to Make a Pizza: Learning a Compositional Layer-Based GAN Model
Dim P. Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, Antonio Torralba; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8002-8011
A food recipe is an ordered set of instructions for preparing a particular dish. From a visual perspective, every instruction step can be seen as a way to change the visual appearance of the dish by adding extra objects (e.g., adding an ingredient) or changing the appearance of the existing ones (e.g., cooking the dish). In this paper, we aim to teach a machine how to make a pizza by building a generative model that mirrors this step-by-step procedure. To do so, we learn composable module operations which are able to either add or remove a particular ingredient. Each operator is designed as a Generative Adversarial Network (GAN). Given only weak image-level supervision, the operators are trained to generate a visual layer that needs to be added to or removed from the existing image. The proposed model is able to decompose an image into an ordered sequence of layers by applying sequentially in the right order the corresponding removing modules. Experimental results on synthetic and real pizza images demonstrate that our proposed model is able to: (1) segment pizza toppings in a weakly- supervised fashion, (2) remove them by revealing what is occluded underneath them (i.e., inpainting), and (3) infer the ordering of the toppings without any depth ordering supervision. Code, data, and models are available online.
****

TransGaGa: Geometry-Aware Unsupervised Image-To-Image Translation
Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8012-8021
Unsupervised image-to-image translation aims at learning a mapping between two visual domains. However, learning a translation across large geometry variations al- ways ends up with failure. In this work, we present a novel disentangle-and-translate framework to tackle the complex objects image-to-image translation task. Instead of learning the mapping on the image space directly, we disentangle image space into a Cartesian product of the appearance and the geometry latent spaces. Specifically, we first in- troduce a geometry prior loss and a conditional VAE loss to encourage the network to learn independent but com- plementary representations. The translation is then built on appearance and geometry space separately. Extensive experiments demonstrate the superior performance of our method to other state-of-the-art approaches, especially in the challenging near-rigid and non-rigid objects translation tasks. In addition, by taking different exemplars as the ap- pearance references, our method also supports multimodal translation. Project page: https://wywu.github. io/projects/TGaGa/TGaGa.html
****

Depth-Attentional Features for Single-Image Rain Removal
Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Pheng-Ann Heng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8022-8031
Rain is a common weather phenomenon, where object visibility varies with depth from the camera and objects faraway are visually blocked more by fog than by rain streaks. Existing methods and datasets for rain removal, however, ignore these physical properties, thereby limiting the rain removal efficiency on real photos. In this work, we first analyze the visual effects of rain subject to scene depth and formulate a rain imaging model collectively with rain streaks and fog; by then, we prepare a new dataset called RainCityscapes with rain streaks and fog on real outdoor photos. Furthermore, we design an end-to-end deep neural network, where we train it to learn depth-attentional features via a depth-guided attention mechanism, and regress a residual map to produce the rain-free image output. We performed various experiments to visually and quantitatively compare our method with several state-of-the-art methods to demonstrate its superiority over t

he others.
****
Hyperspectral Image Reconstruction Using a Deep Spatial-Spectral Prior
Lizhi Wang, Chen Sun, Ying Fu, Min H. Kim, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8032-8041
Regularization is a fundamental technique to solve an ill-posed optimization problem robustly and is essential to reconstruct compressive hyperspectral images. Various hand-crafted priors have been employed as a regularizer but are often insufficient to handle the wide variety of spectra of natural hyperspectral images, resulting in poor reconstruction quality. Moreover, the prior-regularized optimization requires manual tweaking of its weight parameters to achieve a balance between the spatial and spectral fidelity of result images. In this paper, we present a novel hyperspectral image reconstruction algorithm that substitutes the traditional hand-crafted prior with a data-driven prior, based on an optimization-inspired network. Our method consists of two main parts: First, we learn a novel data-driven prior that regularizes the optimization problem with a goal to boost the spatial-spectral fidelity. Our data-driven prior learns both local coherence and dynamic characteristics of natural hyperspectral images. Second, we combine our regularizer with an optimization-inspired network to overcome the heavy computation problem in the traditional iterative optimization methods. We learn the complete parameters in the network through end-to-end training, enabling robust performance with high accuracy. Extensive simulation and hardware experiments validate the superior performance of our method over the state-of-the-art methods.
****
LiFF: Light Field Features in Scale and Depth
Donald G. Dansereau, Bernd Girod, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8042-8051
Feature detectors and descriptors are key low-level vision tools that many higher-level tasks build on. Unfortunately these fail in the presence of challenging light transport effects including partial occlusion, low contrast, and reflective or refractive surfaces. Building on spatio-angular imaging modalities offered by emerging light field cameras, we introduce a new and computationally efficient 4D light field feature detector and descriptor: LiFF. LiFF is scale invariant and utilizes the full 4D light field to detect features that are robust to changes in perspective. This is particularly useful for structure from motion (SfM) and other tasks that match features across viewpoints of a scene. We demonstrate significantly improved 3D reconstructions via SfM when using LiFF instead of the leading 2D or 4D features, and show that LiFF runs an order of magnitude faster than the leading 4D approach. Finally, LiFF inherently estimates depth for each feature, opening a path for future research in light field-based SfM.
****
Deep Exemplar-Based Video Colorization
Bo Zhang, Mingming He, Jing Liao, Pedro V. Sander, Lu Yuan, Amine Bermak, Dong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8052-8061
This paper presents the first end-to-end network for exemplar-based video colorization. The main challenge is to achieve temporal consistency while remaining faithful to the reference style. To address this issue, we introduce a recurrent framework that unifies the semantic correspondence and color propagation steps. Both steps allow a provided reference image to guide the colorization of every frame, thus reducing accumulated propagation errors. Video frames are colorized in sequence based on the colorization history, and its coherency is further enforced by the temporal consistency loss. All of these components, learned end-to-end, help produce realistic videos with good temporal stability. Experiments show our result is superior to the state-of-the-art methods both quantitatively and qualitatively.
****

On Finding Gray Pixels
Yanlin Qian, Joni-Kristian Kamarainen, Jarno Nikkanen, Jiri Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8062-8070
We propose a novel grayness index for finding gray pixels and demonstrate its effectiveness and efficiency in illumination estimation. The grayness index, GI in short, is derived using the Dichromatic Reflection Model and is learning-free. GI allows to estimate one or multiple illumination sources in color-biased images. On standard single-illumination and multiple-illumination estimation benchmarks, GI outperforms state-of-the-art statistical methods and many recent deep methods. GI is simple and fast, written in a few dozen lines of code, processing a 1080p image in 0.4 seconds with a non-optimized Matlab code.
****

UnOS: Unified Unsupervised Optical-Flow and Stereo-Depth Estimation by Watching Videos
Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, Wei Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8071-8081
In this paper, we propose UnOS, an unified system for unsupervised optical flow and stereo depth estimation using convolutional neural network (CNN) by taking advantages of their inherent geometrical consistency based on the rigid-scene assumption. UnOS significantly outperforms other state-of-the-art (SOTA) unsupervised approaches that treated the two tasks independently. Specifically, given two consecutive stereo image pairs from a video, UnOS estimates per-pixel stereo depth images, camera ego-motion and optical flow with three parallel CNNs. Based on these quantities, UnOS computes rigid optical flow and compares it against the optical flow estimated from the FlowNet, yielding pixels satisfying the rigid-scene assumption. Then, we encourage geometrical consistency between the two estimated flows within rigid regions, from which we derive a rigid-aware direct visual odometry (RDVO) module. We also propose rigid and occlusion-aware flow-consistency losses for the learning of UnOS. We evaluated our results on the popular KITTI dataset over 4 related tasks, i.e. stereo depth, optical flow, visual odometry and motion segmentation.
****

Learning Transformation Synchronization
Xiangru Huang, Zhenxiao Liang, Xiaowei Zhou, Yao Xie, Leonidas J. Guibas, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8082-8091
Reconstructing the 3D model of a physical object typically requires us to align the depth scans obtained from different camera poses into the same coordinate system. Solutions to this global alignment problem usually proceed in two steps. The first step estimates relative transformations between pairs of scans using an off-the-shelf technique. Due to limited information presented between pairs of scans, the resulting relative transformations are generally noisy. The second step then jointly optimizes the relative transformations among all input depth scans. A natural constraint used in this step is the cycle-consistency constraint, which allows us to prune incorrect relative transformations by detecting inconsistent cycles. The performance of such approaches, however, heavily relies on the quality of the input relative transformations. Instead of merely using the relative transformations as the input to perform transformation synchronization, we propose to use a neural network to learn the weights associated with each relative transformation. Our approach alternates between transformation synchronization using weighted relative transformations and predicting new weights of the input relative transformations using a neural network. We demonstrate the usefulness of this approach across a wide range of datasets.
****

D2-Net: A Trainable CNN for Joint Description and Detection of Local Features
Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, Torsten Sattler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8092-8101

In this work we address the problem of finding reliable pixel-level corresponden ces under difficult imaging conditions. We propose an approach where a single co nvolutional neural network plays a dual role: It is simultaneously a dense featu re descriptor and a feature detector. By postponing the detection to a later sta ge, the obtained keypoints are more stable than their traditional counterparts b ased on early detection of low-level structures. We show that this model can be trained using pixel correspondences extracted from readily available large-scale SfM reconstructions, without any further annotations. The proposed method obtai ns state-of-the-art performance on both the difficult Aachen Day-Night localizat ion dataset and the InLoc indoor localization benchmark, as well as competitive performance on other benchmarks for image matching and 3D reconstruction.
****

Recurrent Neural Networks With Intra-Frame Iterations for Video Deblurring
Seungjun Nah, Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8102-8111
Recurrent neural networks (RNNs) are widely used for sequential data processing. Recent state-of-the-art video deblurring methods bank on convolutional recurren t neural network architectures to exploit the temporal relationship between neig hboring frames. In this work, we aim to improve the accuracy of recurrent models by adapting the hidden states transferred from past frames to the frame being p rocessed so that the relations between video frames could be better used. We ite ratively update the hidden state via re-using RNN cell parameters before predict ing an output deblurred frame. Since we use existing parameters to update the hi dden state, our method improves accuracy without additional modules. As the arch itecture remains the same regardless of iteration number, fewer iteration models can be considered as a partial computational path of the models with more itera tions. To take advantage of this property, we employ a stochastic method to opti mize our iterative models better. At training time, we randomly choose the itera tion number on the fly and apply a regularization loss that favors less computat ion unless there are considerable reconstruction gains. We show that our method exhibits state-of-the-art video deblurring performance while operating in real-t ime speed.
****

Learning to Extract Flawless Slow Motion From Blurry Videos
Meiguang Jin, Zhe Hu, Paolo Favaro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8112-8121
In this paper, we introduce the task of generating a sharp slow-motion video giv en a low frame rate blurry video. We propose a data-driven approach, where the training data is captured with a high frame rate camera and blurry images are si mulated through an averaging process. While it is possible to train a neural ne twork to recover the sharp frames from their average, there is no guarantee of t he temporal smoothness for the formed video, as the frames are estimated indepen dently. To address the temporal smoothness requirement we propose a system with two networks: One, DeblurNet, to predict sharp keyframes and the second, Interp Net, to predict intermediate frames between the generated keyframes. A smooth tr ansition is ensured by interpolating between consecutive keyframes using InterpN et. Moreover, the proposed scheme enables further increase in frame rate withou t retraining the network, by applying InterpNet recursively between pairs of sha rp frames. We evaluate the proposed method on several datasets, including a nove l dataset captured with a Sony RX V camera. We also demonstrate its performance of increasing the frame rate up to 20 times on real blurry videos.
****

Natural and Realistic Single Image Super-Resolution With Explicit Natural Manifo ld Discrimination
Jae Woong Soh, Gu Yong Park, Junho Jo, Nam Ik Cho; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8122-8 131
Recently, many convolutional neural networks for single image super-resolution (SISR) have been proposed, which focus on reconstructing the high-resolution imag es in terms of objective distortion measures. However, the networks trained with

objective loss functions generally fail to reconstruct the realistic fine textures and details that are essential for better perceptual quality. Recovering the realistic details remains a challenging problem, and only a few works have been proposed which aim at increasing the perceptual quality by generating enhanced textures. However, the generated fake details often make undesirable artifacts and the overall image looks somewhat unnatural. Therefore, in this paper, we present a new approach to reconstructing realistic super-resolved images with high perceptual quality, while maintaining the naturalness of the result. In particular, we focus on the domain prior properties of SISR problem. Specifically, we define the naturalness prior in the low-level domain and constrain the output image in the natural manifold, which eventually generates more natural and realistic images. Our results show better naturalness compared to the recent super-resolution algorithms including perception-oriented ones.

****

RF-Net: An End-To-End Image Matching Network Based on Receptive Field

Xuelun Shen, Cheng Wang, Xin Li, Zenglei Yu, Jonathan Li, Chenglu Wen, Ming Cheng, Zijian He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8132-8140

This paper proposes a new end-to-end trainable matching network based on receptive field, RF-Net, to compute sparse correspondence between images. Building end-to-end trainable matching framework is desirable and challenging. The very recent approach, LF-Net, successfully embeds the entire feature extraction pipeline into a jointly trainable pipeline, and produces the state-of-the-art matching results. This paper introduces two modifications to the structure of LF-Net. First, we propose to construct receptive feature maps, which lead to more effective keypoint detection. Second, we introduce a general loss function term, neighbor mask, to facilitate training patch selection. This results in improved stability in descriptor training. We trained RF-Net on the open dataset HPatches, and compared it with other methods on multiple benchmark datasets. Experiments show that RF-Net outperforms existing state-of-the-art methods.

****

Fast Single Image Reflection Suppression via Convex Optimization

Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, Weiyu Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8141-8149

Removing undesired reflections from images taken through the glass is of great importance in computer vision. It serves as a means to enhance the image quality for aesthetic purposes as well as to preprocess images in machine learning and pattern recognition applications. We propose a convex model to suppress the reflection from a single input image. Our model implies a partial differential equation with gradient thresholding, which is solved efficiently using Discrete Cosine Transform. Extensive experiments on synthetic and real-world images demonstrate that our approach achieves desirable reflection suppression results and dramatically reduces the execution time compared to the state of the art.

****

A Mutual Learning Method for Salient Object Detection With Intertwined Multi-Supervision

Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, Errui Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8150-8159

Though deep learning techniques have made great progress in salient object detection recently, the predicted saliency maps still suffer from incomplete predictions due to the internal complexity of objects and inaccurate boundaries caused by strides in convolution and pooling operations. To alleviate these issues, we propose to train saliency detection networks by exploiting the supervision from not only salient object detection, but also foreground contour detection and edge detection. First, we leverage salient object detection and foreground contour detection tasks in an intertwined manner to generate saliency maps with uniform highlight. Second, the foreground contour and edge detection tasks guide each other simultaneously, thereby leading to preciser foreground contour prediction and

reducing the local noises for edge prediction. In addition, we develop a novel mutual learning module (MLM) which serves as the building block of our method. Each MLM consists of multiple network branches trained in a mutual learning manner, which improves the performance by a large margin. Extensive experiments on seven challenging datasets demonstrate that the proposed method has delivered state-of-the-art results in both salient object detection and edge detection.

****

Enhanced Pix2pix Dehazing Network

Yanyun Qu, Yizi Chen, Jingying Huang, Yuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8160-8168

In this paper, we reduce the image dehazing problem to an image-to-image translation problem, and propose Enhanced Pix2pix Dehazing Network (EPDN), which generates a haze-free image without relying on the physical scattering model. EPDN is embedded by a generative adversarial network, which is followed by a well-designed enhancer. Inspired by visual perception global-first theory, the discriminator guides the generator to create a pseudo realistic image on a coarse scale, while the enhancer following the generator is required to produce a realistic dehazing image on the fine scale. The enhancer contains two enhancing blocks based on the receptive field model, which reinforces the dehazing effect in both color and details. The embedded GAN is jointly trained with the enhancer. Extensive experiment results on synthetic datasets and real-world datasets show that the proposed EPDN is superior to the state-of-the-art methods in terms of PSNR, SSIM, PI, and subjective visual effect.

****

Assessing Personally Perceived Image Quality via Image Features and Collaborative Filtering

Jari Korhonen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8169-8177

During the past few years, different methods for optimizing the camera settings and post-processing techniques to improve the subjective quality of consumer photos have been studied extensively. However, most of the research in the prior art has focused on finding the optimal method for an average user. Since there is large deviation in personal opinions and aesthetic standards, the next challenge is to find the settings and post-processing techniques that fit to the individual users' personal taste. In this study, we aim to predict the personally perceived image quality by combining classical image feature analysis and collaboration filtering approach known from the recommendation systems. The experimental results for the proposed method show promising results. As a practical application, our work can be used for personalizing the camera settings or post-processing parameters for different users and images.

****

Single Image Reflection Removal Exploiting Misaligned Training Data and Network Enhancements

Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, Hua Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8178-8187

Removing undesirable reflections from a single image captured through a glass window is of practical importance to visual computing systems. Although state-of-the-art methods can obtain decent results in certain situations, performance declines significantly when tackling more general real-world cases. These failures stem from the intrinsic difficulty of single image reflection removal -- the fundamental ill-posedness of the problem, and the insufficiency of densely-labeled training data needed for resolving this ambiguity within learning-based neural network pipelines. In this paper, we address these issues by exploiting targeted network enhancements and the novel use of misaligned data. For the former, we augment a baseline network architecture by embedding context encoding modules that are capable of leveraging high-level contextual clues to reduce indeterminacy within areas containing strong reflections. For the latter, we introduce an alignment-invariant loss function that facilitates exploiting misaligned real-world training data that is much easier to collect. Experimental results collectively

show that our method outperforms the state-of-the-art with aligned data, and tha t significant improvements are possible when using additional misaligned data.
****

Exploring Context and Visual Pattern of Relationship for Scene Graph Generation
Wenbin Wang, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 818 8-8197
Relationship is the core of scene graph, but its prediction is far from satisfyi ng because of its complex visual diversity. To alleviate this problem, we treat relationship as an abstract object, exploring not only significative visual patt ern but contextual information for it, which are two key aspects when considerin g object recognition. Our observation on current datasets reveals that there exi sts intimate association among relationships. Therefore, inspired by the success ful application of context to object-oriented tasks, we especially construct con text for relationships where all of them are gathered so that the recognition co uld benefit from their association. Moreover, accurate recognition needs discrim inative visual pattern for object, and so does relationship. In order to discove r effective pattern for relationship, traditional relationship feature extractio n methods such as using union region or combination of subject-object feature pa irs are replaced with our proposed intersection region which focuses on more ess ential parts. Therefore, we present our so-called Relationship Context - InterSe Ction Region (CISC) method. Experiments for scene graph generation on Visual Gen ome dataset and visual relationship prediction on VRD dataset indicate that both the relationship context and intersection region improve performances and reali ze anticipated functions.
****

Learning From Synthetic Data for Crowd Counting in the Wild
Qi Wang, Junyu Gao, Wei Lin, Yuan Yuan; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8198-8207
Recently, counting the number of people for crowd scenes is a hot topic because of its widespread applications (e.g. video surveillance, public security). It is a difficult task in the wild: changeable environment, large-range number of peo ple cause the current methods can not work well. In addition, due to the scarce data, many methods suffer from over-fitting to a different extent. To remedy the above two problems, firstly, we develop a data collector and labeler, which can generate the synthetic crowd scenes and simultaneously annotate them without an y manpower. Based on it, we build a large-scale, diverse synthetic dataset. Seco ndly, we propose two schemes that exploit the synthetic data to boost the perfor mance of crowd counting in the wild: 1) pretrain a crowd counter on the syntheti c data, then finetune it using the real data, which significantly prompts the mo del's performance on real data; 2) propose a crowd counting method via domain ad aptation, which can free humans from heavy data annotations. Extensive experimen ts show that the first method achieves the state-of-the-art performance on four real datasets, and the second outperforms our baselines. The dataset and source code are available at https://gjy3035.github.io/GCC-CL/.
****

A Local Block Coordinate Descent Algorithm for the CSC Model
Ev Zisselman, Jeremias Sulam, Michael Elad; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8208-8217
The Convolutional Sparse Coding (CSC) model has recently gained considerable tra ction in the signal and image processing communities. By providing a global, yet tractable, model that operates on the whole image, the CSC was shown to overcom e several limitations of the patch-based sparse model while achieving superior p erformance in various applications. Contemporary methods for pursuit and learnin g the CSC dictionary often rely on the Alternating Direction Method of Multiplie rs (ADMM) in the Fourier domain for the computational convenience of convolution s, while ignoring the local characterizations of the image. In this work we prop ose a new and simple approach that adopts a localized strategy, based on the Blo ck Coordinate Descent algorithm. The proposed method, termed Local Block Coordin ate Descent (LoBCoD), operates locally on image patches. Furthermore, we introdu

ce a novel stochastic gradient descent version of LoBCoD for training the convol
utional filters. This Stochastic-LoBCoD leverages the benefits of online learnin
g, while being applicable even to a single training image. We demonstrate the ad
vantages of the proposed algorithms for image inpainting and multi-focus image f
usion, achieving state-of-the-art results.
****

Not Using the Car to See the Sidewalk -- Quantifying and Controlling the Effects
 of Context in Classification and Segmentation
Rakshith Shetty, Bernt Schiele, Mario Fritz; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8218-8226
Importance of visual context in scene understanding tasks is well recognized in
 the computer vision community. However, to what extent the computer vision mode
ls are dependent on the context to make their predictions is unclear. A model ov
erly relying on context will fail when encountering objects in different context
s than in training data and hence it is important to identify these dependencies
 before we can deploy the models in the real-world. We propose a method to quant
ify the sensitivity of black-box vision models to visual context by editing imag
es to remove selected objects and measuring the response of the target models. W
e apply this methodology on two tasks, image classification and semantic segment
ation, and discover undesirable dependency between objects and context, for exam
ple that "sidewalk" segmentation is very sensitive to the presence of "cars" in
the image. We propose an object removal based data augmentation solution to miti
gate this dependency and increase the robustness of classification and segmentat
ion models to contextual variations. Our experiments show that the proposed data
 augmentation helps these models improve the performance in out-of-context scena
rios, while preserving the performance on regular data.
****

Discovering Fair Representations in the Data Domain
Novi Quadrianto, Viktoriia Sharmanska, Oliver Thomas; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8227
-8236
Interpretability and fairness are critical in computer vision and machine learni
ng applications, in particular when dealing with human outcomes, e.g. inviting o
r not inviting for a job interview based on application materials that may inclu
de photographs. One promising direction to achieve fairness is by learning data
representations that remove the semantics of protected characteristics, and are
therefore able to mitigate unfair outcomes. All available models however learn l
atent embeddings which comes at the cost of being uninterpretable. We propose to
 cast this problem as data-to-data translation, i.e. learning a mapping from an
input domain to a fair target domain, where a fairness definition is being enfor
ced. Here the data domain can be images, or any tabular data representation. Thi
s task would be straightforward if we had fair target data available, but this i
s not the case. To overcome this, we learn a highly unconstrained mapping by exp
loiting statistics of residuals -- the difference between input data and its tra
nslated version -- and the protected characteristics. When applied to the CelebA
 dataset of face images with gender attribute as the protected characteristic, o
ur model enforces equality of opportunity by adjusting the eyes and lips regions
. Intriguingly, on the same dataset we arrive at similar conclusions when using
semantic attribute representations of images for translation. On face images of
the recent DiF dataset, with the same gender attribute, our method adjusts nose
regions. In the Adult income dataset, also with protected gender attribute, our
model achieves equality of opportunity by, among others, obfuscating the wife an
d husband relationship. Analyzing those systematic changes will allow us to scru
tinize the interplay of fairness criterion, chosen protected characteristics, an
d prediction performance.
****

Actor-Critic Instance Segmentation
Nikita Araslanov, Constantin A. Rothkopf, Stefan Roth; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 823
7-8246

Most approaches to visual scene analysis have emphasised parallel processing of the image elements. However, one area in which the sequential nature of vision is apparent, is that of segmenting multiple, potentially similar and partially occluded objects in a scene. In this work, we revisit the recurrent formulation of this challenging problem in the context of reinforcement learning. Motivated by the limitations of the global max-matching assignment of the ground-truth segments to the recurrent states, we develop an actor-critic approach in which the actor recurrently predicts one instance mask at a time and utilises the gradient from a concurrently trained critic network. We formulate the state, action, and the reward such as to let the critic model long-term effects of the current prediction and in- corporate this information into the gradient signal. Furthermore, to enable effective exploration in the inherently high-dimensional action space of instance masks, we learn a compact representation using a conditional variational auto-encoder. We show that our actor-critic model consistently provides accuracy benefits over the recurrent baseline on standard instance segmentation benchmarks.

****

Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders
Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8247-8255

Many approaches in generalized zero-shot learning rely on cross-modal mapping between the image feature space and the class embedding space. As labeled images are expensive, one direction is to augment the dataset by generating either images or image features. However, the former misses fine-grained details and the latter requires learning a mapping associated with class embeddings. In this work, we take feature generation one step further and propose a model where a shared latent space of image features and class embeddings is learned by modality-specific aligned variational autoencoders. This leaves us with the required discriminative information about the image and classes in the latent features, on which we train a softmax classifier. The key to our approach is that we align the distributions learned from images and from side-information to construct latent features that contain the essential multi-modal information associated with unseen classes. We evaluate our learned latent features on several benchmark datasets, i.e. CUB, SUN, AWA1 and AWA2, and establish a new state of the art on generalized zero-shot as well as on few-shot learning. Moreover, our results on ImageNet with various zero-shot splits show that our latent features generalize well in large-scale settings.

****

Semantic Projection Network for Zero- and Few-Label Semantic Segmentation
Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, Zeynep Akata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8256-8265

Semantic segmentation is one of the most fundamental problems in computer vision and pixel-level labelling in this context is particularly expensive. Hence, there have been several attempts to reduce the annotation effort such as learning from image level labels and bounding box annotations. In this paper we take this one step further and focus on the challenging task of zero- and few-shot learning of semantic segmentation. We define this task as image segmentation by assigning a label to every pixel even though either no labeled sample of that class was present during training, i.e. zero-label semantic segmentation, or only a few labeled samples were present, i.e. few-label semantic segmentation.Our goal is to transfer the knowledge from previously seen classes to novel classes. Our proposed semantic projection network (SPNet) achieves this goal by incorporating a class-level semantic information into any network designed for semantic segmentation, in an end-to-end manner. We also propose a benchmark for this task on the challenging COCO-Stuff and PASCAL VOC12 datasets. Our model is effective in segmenting novel classes, i.e. alleviating expensive dense annotations, but also in adapting to novel classes without forgetting its prior knowledge, i.e. generalized zero- and few-label semantic segmentation.

****
GCAN: Graph Convolutional Adversarial Network for Unsupervised Domain Adaptation
Xinhong Ma, Tianzhu Zhang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8266-8276
To bridge source and target domains for domain adaptation, there are three important types of information including data structure, domain label, and class label. Most existing domain adaptation approaches exploit only one or two types of this information and cannot make them complement and enhance each other. Different from existing methods, we propose an end-to-end Graph Convolutional Adversarial Network (GCAN) for unsupervised domain adaptation by jointly modeling data structure, domain label, and class label in a unified deep framework. The proposed GCAN model enjoys several merits. First, to the best of our knowledge, this is the first work to model the three kinds of information jointly in a deep model for unsupervised domain adaptation. Second, the proposed model has designed three effective alignment mechanisms including structure-aware alignment, domain alignment, and class centroid alignment, which can learn domain-invariant and semantic representations effectively to reduce the domain discrepancy for domain adaptation. Extensive experimental results on five standard benchmarks demonstrate that the proposed GCAN algorithm performs favorably against state-of-the-art unsupervised domain adaptation methods.
****
Seamless Scene Segmentation
Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8277-8286
In this work we introduce a novel, CNN-based architecture that can be trained end-to-end to deliver seamless scene segmentation results. Our goal is to predict consistent semantic segmentation and detection results by means of a panoptic output format, going beyond the simple combination of independently trained segmentation and detection models. The proposed architecture takes advantage of a novel segmentation head that seamlessly integrates multi-scale features generated by a Feature Pyramid Network with contextual information conveyed by a light-weight DeepLab-like module. As additional contribution we review the panoptic metric and propose an alternative that overcomes its limitations when evaluating non-instance categories. Our proposed network architecture yields state-of-the-art results on three challenging street-level datasets, i.e. Cityscapes, Indian Driving Dataset and Mapillary Vistas.
****
Unsupervised Image Matching and Object Discovery as Optimization
Huy V. Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Perez, Jean Ponce; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8287-8296
Learning with complete or partial supervision is power- ful but relies on ever-growing human annotation efforts. As a way to mitigate this serious problem, as well as to serve specific applications, unsupervised learning has emerged as an important field of research. In computer vision, unsu- pervised learning comes in various guises. We focus here on the unsupervised discovery and matching of object cate- gories among images in a collection, following the work of Cho et al. [12]. We show that the original approach can be reformulated and solved as a proper optimization problem. Experiments on several benchmarks establish the merit of our approach.
****
Wide-Area Crowd Counting via Ground-Plane Density Maps and Multi-View Fusion CNNs
Qi Zhang, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8297-8306
Crowd counting in single-view images has achieved outstanding performance on existing counting datasets. However, single-view counting is not applicable to large and wide scenes (e.g., public parks, long subway platforms, or event spaces) because a single camera cannot capture the whole scene in adequate detail for cou

nting, e.g., when the scene is too large to fit into the field-of-view of the ca
mera, too long so that the resolution is too low on faraway crowds, or when the
re are too many large objects that occlude large portions of the crowd. Therefor
e, to solve the wide-area counting task requires multiple cameras with overlappi
ng fields-of-view. In this paper, we propose a deep neural network framework fo
r multi-view crowd counting, which fuses information from multiple camera views
to predict a scene-level density map on the ground-plane of the 3D world. We con
sider 3 versions of the fusion framework: the late fusion model fuses camera-vie
w density map; the naive early fusion model fuses camera-view feature maps; and
the multi-view multi-scale early fusion model favors that features aligned to th
e same ground-plane point have consistent scales. We test our 3 fusion models on
 3 multi-view counting datasets, PETS2009, DukeMTMC, and a newly collected multi
-view counting dataset containing a crowded street intersection. Our methods ach
ieve state-of-the-art results compared to other multi-view counting baselines.
****

Show, Control and Tell: A Framework for Generating Controllable and Grounded Cap
tions

Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8307-831
6

Current captioning approaches can describe images using black-box architectures
whose behavior is hardly controllable and explainable from the exterior. As an i
mage can be described in infinite ways depending on the goal and the context at
hand, a higher degree of controllability is needed to apply captioning algorithm
s in complex scenarios. In this paper, we introduce a novel framework for image
captioning which can generate diverse descriptions by allowing both grounding an
d controllability. Given a control signal in the form of a sequence or set of im
age regions, we generate the corresponding caption through a recurrent architect
ure which predicts textual chunks explicitly grounded on regions, following the
constraints of the given control. Experiments are conducted on Flickr30k Entitie
s and on COCO Entities, an extended version of COCO in which we add grounding an
notations collected in a semi-automatic manner. Results demonstrate that our met
hod achieves state of the art performances on controllable image captioning, in
terms of caption quality and diversity. Code and annotations are publicly availa
ble at: https://github.com/aimagelab/show-control-and-tell.
****

Towards VQA Models That Can Read

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv B
atra, Devi Parikh, Marcus Rohrbach; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8317-8326

Studies have shown that a dominant class of questions asked by visually impaired
 users on images of their surroundings involves reading text in the image. But t
oday's VQA models can not read! Our paper takes a first step towards addressing
this problem. First, we introduce a new "TextVQA" dataset to facilitate progress
 on this important problem. Existing datasets either have a small proportion of
questions about text (e.g., the VQA dataset) or are too small (e.g., the VizWiz
dataset). TextVQA contains 45,336 questions on 28,408 images that require reason
ing about text to answer. Second, we introduce a novel model architecture that r
eads text in the image, reasons about it in the context of the image and the que
stion, and predicts an answer which might be a deduction based on the text and t
he image or composed of the strings found in the image. Consequently, we call ou
r approach Look, Read, Reason & Answer (LoRRA). We show that LoRRA outperforms e
xisting state-of-the-art VQA models on our TextVQA dataset. We find that the gap
 between human performance and machine performance is significantly larger on Te
xtVQA than on VQA 2.0, suggesting that TextVQA is well-suited to benchmark progr
ess along directions complementary to VQA 2.0.
****

Object-Aware Aggregation With Bidirectional Temporal Graph for Video Captioning

Junchao Zhang, Yuxin Peng; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2019, pp. 8327-8336

Video captioning aims to automatically generate natural language descriptions of video content, which has drawn a lot of attention recent years. Generating accurate and fine-grained captions needs to not only understand the global content of video, but also capture the detailed object information. Meanwhile, video representations have great impact on the quality of generated captions. Thus, it is important for video captioning to capture salient objects with their detailed temporal dynamics, and represent them using discriminative spatio-temporal representations. In this paper, we propose a new video captioning approach based on object-aware aggregation with bidirectional temporal graph (OA-BTG), which captures detailed temporal dynamics for salient objects in video, and learns discriminative spatio-temporal representations by performing object-aware local feature aggregation on detected object regions. The main novelties and advantages are: (1) Bidirectional temporal graph: A bidirectional temporal graph is constructed along and reversely along the temporal order, which provides complementary ways to capture the temporal trajectories for each salient object. (2) Object-aware aggregation: Learnable VLAD (Vector of Locally Aggregated Descriptors) models are constructed on object temporal trajectories and global frame sequence, which performs object-aware aggregation to learn discriminative representations. A hierarchical attention mechanism is also developed to distinguish different contributions of multiple objects. Experiments on two widely-used datasets demonstrate our OA-BTG achieves state-of-the-art performance in terms of BLEU@4, METEOR and CIDEr metrics.

****

Progressive Attention Memory Network for Movie Story Question Answering

Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, Chang D. Yoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8337-8346

This paper proposes the progressive attention memory network (PAMN) for movie story question answering (QA). Movie story QA is challenging compared to VQA in two aspects: (1) pinpointing the temporal parts relevant to answer the question is difficult as the movies are typically longer than an hour, (2) it has both video and subtitle where different questions require different modality to infer the answer. To overcome these challenges, PAMN involves three main features: (1) progressive attention mechanism that utilizes cues from both question and answer to progressively prune out irrelevant temporal parts in memory, (2) dynamic modality fusion that adaptively determines the contribution of each modality for answering the current question, and (3) belief correction answering scheme that successively corrects the prediction score on each candidate answer. Experiments on publicly available benchmark datasets, MovieQA and TVQA, demonstrate that each feature contributes to our movie story QA architecture, PAMN, and improves performance to achieve the state-of-the-art result. Qualitative analysis by visualizing the inference mechanism of PAMN is also provided.

****

Memory-Attended Recurrent Network for Video Captioning

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8347-8356

Typical techniques for video captioning follow the encoder-decoder framework, which can only focus on one source video being processed. A potential disadvantage of such design is that it cannot capture the multiple visual context information of a word appearing in more than one relevant videos in training data. To tackle this limitation, we propose the Memory-Attended Recurrent Network (MARN) for video captioning, in which a memory structure is designed to explore the full-spectrum correspondence between a word and its various similar visual contexts across videos in training data. Thus, our model is able to achieve a more comprehensive understanding for each word and yield higher captioning quality. Furthermore, the built memory structure enables our method to model the compatibility between adjacent words explicitly instead of asking the model to learn implicitly, as most existing models do. Extensive validation on two real-word datasets demonstrates that our MARN consistently outperforms state-of-the-art methods.

****
Visual Query Answering by Entity-Attribute Graph Matching and Reasoning
Peixi Xiong, Huayi Zhan, Xin Wang, Baivab Sinha, Ying Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8357-8366
Visual Query Answering (VQA) is of great significance in offering people convenience: one can raise a question for details of objects, or high-level understanding about the scene, over an image. This paper proposes a novel method to address the VQA problem. In contrast to prior works, our method that targets single scene VQA, replies on graph-based techniques and involves reasoning. In a nutshell, our approach is centered on three graphs. The first graph, referred to as inference graph G_I, is constructed via learning over labeled data. The other two graphs, referred to as query graph Q and entity-attribute graph EAG, are generated from natural language query NLQ and image Img, that are issued from users, respectively. As EAG often does not take sufficient information to answer Q, we develop techniques to infer missing information of EAG with G_I. Based on EAG and Q, we provide techniques to find matches of Q in EAG, as the answer of NLQ in Img. Unlike commonly used VQA methods that are based on end-to-end neural networks, our graph-based method shows well-designed reasoning capability, and thus is highly interpretable. We also create a dataset on soccer match (Soccer-VQA) with rich annotations. The experimental results show that our approach outperforms the state-of-the-art method and has high potential for future investigation.
****
Look Back and Predict Forward in Image Captioning
Yu Qin, Jiajun Du, Yonghua Zhang, Hongtao Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8367-8375
Most existing attention-based methods on image captioning focus on the current word and visual information in one time step and generate the next word, without considering the visual and linguistic coherence. We propose Look Back (LB) method to embed visual information from the past and Predict Forward (PF) approach to look into future. LB method introduces attention value from the previous time step into the current attention generation to suit visual coherence of human. PF model predicts the next two words in one time step and jointly employs their probabilities for inference. Then the two approaches are combined together as LBPF to further integrate visual information from the past and linguistic information in the future to improve image captioning performance. All the three methods are applied on a classic base decoder, and show remarkable improvements on MSCOCO dataset with small increments on parameter counts. Our LBPF model achieves BLEU-4 / CIDEr / SPICE scores of 37.4 / 116.4 / 21.2 with cross-entropy loss and 38.3 / 127.6 / 22.0 with CIDEr optimization. Our three proposed methods can be easily applied on most attention-based encoder-decoder models for image captioning.
****
Explainable and Explicit Visual Reasoning Over Scene Graphs
Jiaxin Shi, Hanwang Zhang, Juanzi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8376-8384
We aim to dismantle the prevalent black-box neural architectures used in complex visual reasoning tasks, into the proposed eXplainable and eXplicit Neural Modules (XNMs), which advance beyond existing neural module networks towards using scene graphs --- objects as nodes and the pairwise relationships as edges --- for explainable and explicit reasoning with structured knowledge. XNMs allow us to pay more attention to teach machines how to "think", regardless of what they "look". As we will show in the paper, by using scene graphs as an inductive bias, 1) we can design XNMs in a concise and flexible fashion, i.e., XNMs merely consist of 4 meta-types, which significantly reduce the number of parameters by 10 to 100 times, and 2) we can explicitly trace the reasoning-flow in terms of graph attentions. XNMs are so generic that they support a wide range of scene graph implementations with various qualities. For example, when the graphs are detected perfectly, XNMs achieve 100% accuracy on both CLEVR and CLEVR CoGenT, establishing an empirical performance upper-bound for visual reasoning; when the graphs are noisily detected from real-world images, XNMs are still robust to achieve a comp

etitive 67.5% accuracy on VQAv2.0, surpassing the popular bag-of-objects attenti
on models without graph structures.
****

Transfer Learning via Unsupervised Task Discovery for Visual Question Answering
Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, Bohyung Han; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 838
5-8394
We study how to leverage off-the-shelf visual and linguistic data to cope with o
ut-of-vocabulary answers in visual question answering task. Existing large-scale
visual datasets with annotations such as image class labels, bounding boxes and
region descriptions are good sources for learning rich and diverse visual conce
pts. However, it is not straightforward how the visual concepts can be captured
and transferred to visual question answering models due to missing link between
question dependent answering models and visual data without question. We tackle
this problem in two steps: 1) learning a task conditional visual classifier, whi
ch is capable of solving diverse question-specific visual recognition tasks, bas
ed on unsupervised task discovery and 2) transferring the task conditional visua
l classifier to visual question answering models. Specifically, we employ lingui
stic knowledge sources such as structured lexical database (e.g. WordNet) and vi
sual descriptions for unsupervised task discovery, and transfer a learned task c
onditional visual classifier as an answering unit in a visual question answering
model. We empirically show that the proposed algorithm generalizes to out-of-vo
cabulary answers successfully using the knowledge transferred from the visual da
taset.
****

Intention Oriented Image Captions With Guiding Objects
Yue Zheng, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2019, pp. 8395-8404
Although existing image caption models can produce promising results using recur
rent neural networks (RNNs), it is difficult to guarantee that an object we care
about is contained in generated descriptions, for example in the case that the
object is inconspicuous in the image. Problems become even harder when these obj
ects did not appear in training stage. In this paper, we propose a novel approac
h for generating image captions with guiding objects (CGO). The CGO constrains t
he model to involve a human-concerned object when the object is in the image. CG
O ensures that the object is in the generated description while maintaining flue
ncy. Instead of generating the sequence from left to right, we start the descrip
tion with a selected object and generate other parts of the sequence based on th
is object. To achieve this, we design a novel framework combining two LSTMs in o
pposite directions. We demonstrate the characteristics of our method on MSCOCO w
here we generate descriptions for each detected object in the images. With CGO,
we can extend the ability of description to the objects being neglected in image
caption labels and provide a set of more comprehensive and diverse descriptions
for an image. CGO shows advantages when applied to the task of describing novel
objects. We show experimental results on both MSCOCO and ImageNet datasets. Eva
luations show that our method outperforms the state-of-the-art models in the tas
k with average F1 75.8, leading to better descriptions in terms of both content
accuracy and fluency.
****

Uncertainty Guided Multi-Scale Residual Learning-Using a Cycle Spinning CNN for
Single Image De-Raining
Rajeev Yasarla, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2019, pp. 8405-8414
Single image de-raining is an extremely challenging problem since the rainy imag
e may contain rain streaks which may vary in size, direction and density. Previo
us approaches have attempted to address this problem by leveraging some prior in
formation to remove rain streaks from a single image. One of the major limitatio
ns of these approaches is that they do not consider the location information of
rain drops in the image. The proposed Uncertainty guided Multi-scale Residual Le
arning (UMRL) network attempts to address this issue by learning the rain conten

t at different scales and using them to estimate the final de-rained output. In addition, we introduce a technique which guides the network to learn the network weights based on the confidence measure about the estimate. Furthermore, we introduce a new training and testing procedure based on the notion of cycle spinning to improve the final de-raining performance. Extensive experiments on synthetic and real datasets to demonstrate that the proposed method achieves significant improvements over the recent state-of-the-art methods.
****

Toward Realistic Image Compositing With Adversarial Learning
Bor-Chun Chen, Andrew Kae; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8415-8424
Compositing a realistic image is a challenging task and usually requires considerable human supervision using professional image editing software. In this work we propose a generative adversarial network (GAN) architecture for automatic image compositing. The proposed model consists of four sub-networks: a transformation network that improves the geometric and color consistency of the composite image, a refinement network that polishes the boundary of the composite image, and a pair of discriminator network and a segmentation network for adversarial learning. Experimental results on both synthesized images and real images show that our model, Geometrically and Color Consistent GANs (GCC-GANs), can automatically generate realistic composite images compared to several state-of-the-art methods, and does not require any manual effort.
****

Cross-Classification Clustering: An Efficient Multi-Object Tracking Technique for 3-D Instance Segmentation in Connectomics
Yaron Meirovitch, Lu Mi, Hayk Saribekyan, Alexander Matveev, David Rolnick, Nir Shavit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8425-8435
Pixel-accurate tracking of objects is a key element in many computer vision applications, often solved by iterated individual object tracking or instance segmentation followed by object matching. Here we introduce cross-classification clustering (3C), a technique that simultaneously tracks complex, interrelated objects in an image stack. The key idea in cross-classification is to efficiently turn a clustering problem into a classification problem by running a logarithmic number of independent classifications per image, letting the cross-labeling of these classifications uniquely classify each pixel to the object labels. We apply the 3C mechanism to achieve state-of-the-art accuracy in connectomics -- the nanoscale mapping of neural tissue from electron microscopy volumes. Our reconstruction system increases scalability by an order of magnitude over existing single-object tracking methods (such as flood-filling networks). This scalability is important for the deployment of connectomics pipelines, since currently the best performing techniques require computing infrastructures that are beyond the reach of most laboratories. Our algorithm may offer benefits in other domains that require pixel-accurate tracking of multiple objects, such as segmentation of videos and medical imagery.
****

Deep ChArUco: Dark ChArUco Marker Pose Estimation
Danying Hu, Daniel DeTone, Tomasz Malisiewicz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8436-8444
ChArUco boards are used for camera calibration, monocular pose estimation, and pose verification in both robotics and augmented reality. Such fiducials are detectable via traditional computer vision methods (as found in OpenCV) in well-lit environments, but classical methods fail when the lighting is poor or when the image undergoes extreme motion blur. We present Deep ChArUco, a real-time pose estimation system which combines two custom deep networks, ChArUcoNet and RefineNet, with the Perspective-n-Point (PnP) algorithm to estimate the marker's 6DoF pose. ChArUcoNet is a two-headed marker-specific convolutional neural network (CNN) which jointly outputs ID-specific classifiers and 2D point locations. The 2D point locations are further refined into subpixel coordinates using RefineNet. Our networks are trained using a combination of auto-labeled videos of the target

marker, synthetic subpixel corner data, and extreme data augmentation. We evaluate Deep ChArUco in challenging low-light, high-motion, high-blur scenarios and demonstrate that our approach is superior to a traditional OpenCV-based method for ChArUco marker detection and pose estimation.
****

Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving
Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8445-8453
3D object detection is an essential task in autonomous driving. Recent techniques excel with highly accurate detection rates, provided the 3D input data is obtained from precise but expensive LiDAR technology. Approaches based on cheaper monocular or stereo imagery data have, until now, resulted in drastically lower accuracies --- a gap that is commonly attributed to poor image-based depth estimation. However, in this paper we argue that it is not the quality of the data but its representation that accounts for the majority of the difference. Taking the inner workings of convolutional neural networks into consideration, we propose to convert image-based depth maps to pseudo-LiDAR representations --- essentially mimicking the LiDAR signal. With this representation we can apply different existing LiDAR-based detection algorithms. On the popular KITTI benchmark, our approach achieves impressive improvements over the existing state-of-the-art in image-based performance --- raising the detection accuracy of objects within the 30 m range from the previous state-of-the-art of 22% to an unprecedented 74%. At the time of submission our algorithm holds the highest entry on the KITTI 3D object detection leaderboard for stereo-image-based approaches.
****

Rules of the Road: Predicting Driving Behavior With a Convolutional Model of Semantic Interactions
Joey Hong, Benjamin Sapp, James Philbin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8454-8462
We focus on the problem of predicting future states of entities in complex, real-world driving scenarios. Previous research has approached this problem via low-level signals to predict short time horizons, and has not addressed how to leverage key assets relied upon heavily by industry self-driving systems: (1) large 3D perception efforts which provide highly accurate 3D states of agents with rich attributes, and (2) detailed and accurate semantic maps of the environment (lanes, traffic lights, crosswalks, etc). We present a unified representation which encodes such high-level semantic information in a spatial grid, allowing the use of deep convolutional models to fuse complex scene context. This enables learning entity-entity and entity-environment interactions with simple, feed-forward computations in each timestep within an overall temporal model of an agent's behavior. We propose different ways of modelling the future as a distribution over future states using standard supervised learning. We introduce a novel dataset providing industry-grade rich perception and semantic inputs, and empirically show we can effectively learn fundamentals of driving behavior.
****

Metric Learning for Image Registration
Marc Niethammer, Roland Kwitt, Francois-Xavier Vialard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8463-8472
Image registration is a key technique in medical image analysis to estimate deformations between image pairs. A good deformation model is important for high-quality estimates. However, most existing approaches use ad-hoc deformation models chosen for mathematical convenience rather than to capture observed data variation. Recent deep learning approaches learn deformation models directly from data. However, they provide limited control over the spatial regularity of transformations. Instead of learning the entire registration approach, we learn a spatially-adaptive regularizer within a registration model. This allows controlling the desired level of regularity and preserving structural properties of a registrati

on model. For example, diffeomorphic transformations can be attained. Our approa ch is a radical departure from existing deep learning approaches to image regist ration by embedding a deep learning model in an optimization-based registration algorithm to parameterize and data-adapt the registration model itself.
****

LO-Net: Deep Real-Time Lidar Odometry
Qing Li, Shaoyang Chen, Cheng Wang, Xin Li, Chenglu Wen, Ming Cheng, Jonat han Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2019, pp. 8473-8482
We present a novel deep convolutional network pipeline, LO-Net, for real-time li dar odometry estimation. Unlike most existing lidar odometry (LO) estimations th at go through individually designed feature selection, feature matching, and pos e estimation pipeline, LO-Net can be trained in an end-to-end manner. With a new mask-weighted geometric constraint loss, LO-Net can effectively learn feature r epresentation for LO estimation, and can implicitly exploit the sequential depen dencies and dynamics in the data. We also design a scan-to-map module, which use s the geometric and semantic information learned in LO-Net, to improve the estim ation accuracy. Experiments on benchmark datasets demonstrate that LO-Net outper forms existing learning based approaches and has similar accuracy with the state -of-the-art geometry-based approach, LOAM.
****

TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions
Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, Dinesh Manocha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 019, pp. 8483-8492
We present a new algorithm for predicting the near-term trajectories of road age nts in dense traffic videos. Our approach is designed for heterogeneous traffic, where the road agents may correspond to buses, cars, scooters, bi-cycles, or pe destrians. We model the interactions between different road agents using a novel LSTM-CNN hybrid network for trajectory prediction. In particular, we take into account heterogeneous interactions that implicitly account for the varying shape s, dynamics, and behaviors of different road agents. In addition, we model horiz on-based interactions which are used to implicitly model the driving behavior of each road agent. We evaluate the performance of our prediction algorithm, TraPH ic, on the standard datasets and also introduce a new dense, heterogeneous traff ic dataset corresponding to urban Asian videos and agent trajectories. We outper form state-of-the-art methods on dense traffic datasets by 30%.
****

World From Blur
Jiayan Qiu, Xinchao Wang, Stephen J. Maybank, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp . 8493-8504
What can we tell from a single motion-blurred image? We show in this paper that a 3D scene can be revealed. Unlike prior methods that focus on producing a deb lurred image, we propose to estimate and take advantage of the hidden message of a blurred image, the relative motion trajectory, to restore the 3D scene collap sed during the exposure process. To this end, we train a deep network that joint ly predicts the motion trajectory, the deblurred image, and the depth one, all o f which in turn form a collaborative and self-supervised cycle that supervise on e another to reproduce the input blurred image, enabling plausible 3D scene rec onstruction from a single blurred image. We test the proposed model on several l arge-scale datasets we constructed based on benchmarks, as well as real-world bl urred images, and show that it yields very encouraging quantitative and qualitat ive results.
****

Topology Reconstruction of Tree-Like Structure in Images via Structural Similari ty Measure and Dominant Set Clustering
Jianyang Xie, Yitian Zhao, Yonghuai Liu, Pan Su, Yifan Zhao, Jun Cheng, Ya lin Zheng, Jiang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR), 2019, pp. 8505-8513
The reconstruction and analysis of tree-like topological structures in the biomedical images is crucial for biologists and surgeons to understand biomedical conditions and plan surgical procedures. The underlying tree-structure topology reveals how different curvilinear components are anatomically connected to each other. Existing automated topology reconstruction methods have great difficulty in identifying the connectivity when two or more curvilinear components cross or bifurcate, due to their projection ambiguity, imaging noise and low contrast. In this paper, we propose a novel curvilinear structural similarity measure to guide a dominant-set clustering approach to address this indispensable issue. The novel similarity measure takes into account both intensity and geometric properties in representing the curvilinear structure locally and globally, and group curvilinear objects at crossover points into different connected branches by dominant-set clustering. The proposed method is applicable to different imaging modalities, and quantitative and qualitative results on retinal vessel, plant root, and neuronal network datasets show that our methodology is capable of advancing the current state-of-the-art techniques.
****

Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training
Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8514-8522
Most existing Re-IDentification (Re-ID) methods are highly dependent on precise bounding boxes that enable images to be aligned with each other. However, due to the challenging practical scenarios, current detection models often produce inaccurate bounding boxes, which inevitably degenerate the performance of existing Re-ID algorithms. In this paper, we propose a novel coarse-to-fine pyramid model to relax the need of bounding boxes, which not only incorporates local and global information, but also integrates the gradual cues between them. The pyramid model is able to match at different scales and then search for the correct image of the same identity, even when the image pairs are not aligned. In addition, in order to learn discriminative identity representation, we explore a dynamic training scheme to seamlessly unify two losses and extract appropriate shared information between them. Experimental results clearly demonstrate that the proposed method achieves the state-of-the-art results on three datasets. Especially, our approach exceeds the current best method by 9.5% on the most challenging CUHK03 dataset.
****