3D-RelNet: Joint Object and Relational Network for 3D Prediction

Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, Abhinav Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2212-2221

We propose an approach to predict the 3D shape and pose for the objects present in a scene. Existing learning based methods that pursue this goal make independent predictions per object, and do not leverage the relationships amongst them. We argue that reasoning about these relationships is crucial, and present an approach to incorporate these in a 3D prediction framework. In addition to independent per-object predictions, we predict pairwise relations in the form of relative 3D pose, and demonstrate that these can be easily incorporated to improve object level estimates. We report performance across different datasets (SUNCG, NYUv2), and show that our approach significantly improves over independent prediction approaches while also outperforming alternate implicit reasoning methods.

****

Sampling-Free Epistemic Uncertainty Estimation Using Approximated Variance Propagation

Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2931-2940

We present a sampling-free approach for computing the epistemic uncertainty of a neural network. Epistemic uncertainty is an important quantity for the deployment of deep neural networks in safety-critical applications, since it represents how much one can trust predictions on new data. Recently promising works were proposed using noise injection combined with Monte-Carlo sampling at inference time to estimate this quantity (e.g. Monte-Carlo dropout). Our main contribution is an approximation of the epistemic uncertainty estimated by these methods that does not require sampling, thus notably reducing the computational overhead. We apply our approach to large-scale visual tasks (i.e., semantic segmentation and depth regression) to demonstrate the advantages of our method compared to sampling-based approaches in terms of quality of the uncertainty estimates as well as of computational overhead.

****

Universal Adversarial Perturbation via Prior Driven Uncertainty Approximation

Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, Feiyue Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2941-2949

Deep learning models have shown their vulnerabilities to universal adversarial perturbations (UAP), which are quasi-imperceptible. Compared to the conventional supervised UAPs that suffer from the knowledge of training data, the data-independent unsupervised UAPs are more applicable. Existing unsupervised methods fail to take advantage of the model uncertainty to produce robust perturbations. In this paper, we propose a new unsupervised universal adversarial perturbation method, termed as Prior Driven Uncertainty Approximation (PD-UA), to generate a robust UAP by fully exploiting the model uncertainty at each network layer. Specifically, a Monte Carlo sampling method is deployed to activate more neurons to increase the model uncertainty for a better adversarial perturbation. Thereafter, a textural bias prior to revealing a statistical uncertainty is proposed, which helps to improve the attacking performance. The UAP is crafted by the stochastic gradient descent algorithm with a boosted momentum optimizer, and a Laplacian pyramid frequency model is finally used to maintain the statistical uncertainty. Extensive experiments demonstrate that our method achieves well attacking performances on the ImageNet validation set, and significantly improves the fooling rate compared with the state-of-the-art methods.

****

Understanding Deep Networks via Extremal Perturbations and Smooth Masks

Ruth Fong, Mandela Patrick, Andrea Vedaldi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2950-2958

Attribution is the problem of finding which parts of an image are the most responsible for the output of a deep neural network. An important family of attributi on methods is based on measuring the effect of perturbations applied to the input image, either via exhaustive search or by finding representative perturbations via optimization. In this paper, we discuss some of the shortcomings of existing approaches to perturbation analysis and address them by introducing the concept of extremal perturbations, which are theoretically grounded and interpretable. We also introduce a number of technical innovations to compute these extremal perturbations, including a new area constraint and a parametric family of smooth perturbations, which allow us to remove all tunable weighing factors from the optimization problem. We analyze the effect of perturbations as a function of their area, demonstrating excellent sensitivity to the spatial properties of the network under stimulation. We also extend perturbation analysis to the intermediate layers of a deep neural network. This application allows us to show how compactly an image can be represented (in terms of the number of channels it requires). We also demonstrate that the consistency with which images of a given class rely on the same intermediate channel correlates well with class accuracy.

****

Unsupervised Pre-Training of Image Features on Non-Curated Data
Mathilde Caron, Piotr Bojanowski, Julien Mairal, Armand Joulin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2959-2968
Pre-training general-purpose visual features with convolutional neural networks without relying on annotations is a challenging and important task. Most recent efforts in unsupervised feature learning have focused on either small or highly curated datasets like ImageNet, whereas using uncurated raw datasets was found to decrease the feature quality when evaluated on a transfer task. Our goal is to bridge the performance gap between unsupervised methods trained on curated data, which are costly to obtain, and massive raw datasets that are easily available. To that effect, we propose a new unsupervised approach which leverages self-supervision and clustering to capture complementary statistics from large-scale data. We validate our approach on 96 million images from YFCC100M, achieving state-of-the-art results among unsupervised methods on standard benchmarks, which confirms the potential of unsupervised learning when only uncurated data are available. We also show that pre-training a supervised VGG-16 with our method achieves 74.9% top-1 classification accuracy on the validation set of ImageNet, which is an improvement of +0.8% over the same network trained from scratch. Our code is available at https://github.com/facebookresearch/DeeperCluster.

****

Learning Local Descriptors With a CDF-Based Dynamic Soft Margin
Linguang Zhang, Szymon Rusinkiewicz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2969-2978
The triplet loss is adopted by a variety of learning tasks, such as local feature descriptor learning. However, its standard formulation with a hard margin only leverages part of the training data in each mini-batch. Moreover, the margin is often empirically chosen or determined through computationally expensive validation, and stays unchanged during the entire training session. In this work, we propose a simple yet effective method to overcome the above limitations. The core idea is to replace the hard margin with a non-parametric soft margin, which is dynamically updated. The major observation is that the difficulty of a triplet can be inferred from the cumulative distribution function of the triplets' signed distances to the decision boundary. We demonstrate through experiments on both real-valued and binary local feature descriptors that our method leads to state-of-the-art performance on popular benchmarks, while eliminating the need to determine the best margin.

****

Bayes-Factor-VAE: Hierarchical Bayesian Deep Auto-Encoder Models for Factor Disentanglement
Minyoung Kim, Yuting Wang, Pritish Sahu, Vladimir Pavlovic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2979-2987
We propose a family of novel hierarchical Bayesian deep auto-encoder models capa

ble of identifying disentangled factors of variability in data. While many recent attempts at factor disentanglement have focused on sophisticated learning objectives within the VAE framework, their choice of a standard normal as the latent factor prior is both suboptimal and detrimental to performance. Our key observation is that the disentangled latent variables responsible for major sources of variability, the relevant factors, can be more appropriately modeled using long-tail distributions. The typical Gaussian priors are, on the other hand, better suited for modeling of nuisance factors. Motivated by this, we extend the VAE to a hierarchical Bayesian model by introducing hyper-priors on the variances of Gaussian latent priors, mimicking an infinite mixture, while maintaining tractable learning and inference of the traditional VAEs. This analysis signifies the importance of partitioning and treating in a different manner the latent dimensions corresponding to relevant factors and nuisances. Our proposed models, dubbed Bayes-Factor-VAEs, are shown to outperform existing methods both quantitatively and qualitatively in terms of latent disentanglement across several challenging benchmark tasks.

****

Linearized Multi-Sampling for Differentiable Image Transformation

Wei Jiang, Weiwei Sun, Andrea Tagliasacchi, Eduard Trulls, Kwang Moo Yi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2988-2997

We propose a novel image sampling method for differentiable image transformation in deep neural networks. The sampling schemes currently used in deep learning, such as Spatial Transformer Networks, rely on bilinear interpolation, which performs poorly under severe scale changes, and more importantly, results in poor gradient propagation. This is due to their strict reliance on direct neighbors. Instead, we propose to generate random auxiliary samples in the vicinity of each pixel in the sampled image, and create a linear approximation with their intensity values. We then use this approximation as a differentiable formula for the transformed image. We demonstrate that our approach produces more representative gradients with a wider basin of convergence for image alignment, which leads to considerable performance improvements when training networks for registration and classification tasks. This is not only true under large downsampling, but also when there are no scale changes. We compare our approach with multi-scale sampling and show that we outperform it. We then demonstrate that our improvements to the sampler are compatible with other tangential improvements to Spatial Transformer Networks and that it further improves their performance.

****

AdaTransform: Adaptive Data Transformation

Zhiqiang Tang, Xi Peng, Tingfeng Li, Yizhe Zhu, Dimitris N. Metaxas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2998-3006

Data augmentation is widely used to increase data variance in training deep neural networks. However, previous methods require either comprehensive domain knowledge or high computational cost. Can we learn data transformation automatically and efficiently with limited domain knowledge? Furthermore, can we leverage data transformation to improve not only network training but also network testing? In this work, we propose adaptive data transformation to achieve the two goals. The AdaTransform can increase data variance in training and decrease data variance in testing. Experiments on different tasks prove that it can improve generalization performance.

****

CARAFE: Content-Aware ReAssembly of FEatures

Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3007-3016

Feature upsampling is a key operation in a number of modern convolutional network architectures, e.g. feature pyramids. Its design is critical for dense prediction tasks such as object detection and semantic/instance segmentation. In this work, we propose Content-Aware ReAssembly of FEatures (CARAFE), a universal, ligh tweight and highly effective operator to fulfill this goal. CARAFE has several appealing properties: (1) Large field of view. Unlike previous works (e.g. bilinear interpolation) that only exploit subpixel neighborhood, CARAFE can aggregate contextual information within a large receptive field. (2) Content-aware handling. Instead of using a fixed kernel for all samples (e.g. deconvolution), CARAFE enables instance-specific content-aware handling, which generates adaptive kernels on-the-fly. (3) Lightweight and fast to compute. CARAFE introduces little computational overhead and can be readily integrated into modern network architectures. We conduct comprehensive evaluations on standard benchmarks in object detection, instance/semantic segmentation and inpainting. CARAFE shows consistent and substantial gains across all the tasks (1.2% AP, 1.3% AP, 1.8% mIoU, 1.1dB respectively) with negligible computational overhead. It has great potential to serve as a strong building block for future research. Code and models are available at https://github.com/open-mmlab/mmdetection.

****

AFD-Net: Aggregated Feature Difference Learning for Cross-Spectral Image Patch Matching

Dou Quan, Xuefeng Liang, Shuang Wang, Shaowei Wei, Yanfeng Li, Ning Huyan, Licheng Jiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3017-3026

Image patch matching across different spectral domains is more challenging than in a single spectral domain. We consider the reason is twofold: 1. the weaker discriminative feature learned by conventional methods; 2. the significant appearance difference between two images domains. To tackle these problems, we propose an aggregated feature difference learning network (AFD-Net). Unlike other methods that merely rely on the high-level features, we find the feature differences in other levels also provide useful learning information. Thus, the multi-level feature differences are aggregated to enhance the discrimination. To make features invariant across different domains, we introduce a domain invariant feature extraction network based on instance normalization (IN). In order to optimize the AFD-Net, we borrow the large margin cosine loss which can minimize intra-class distance and maximize inter-class distance between matching and non-matching samples. Extensive experiments show that AFD-Net largely outperforms the state-of-the-arts on the cross-spectral dataset, meanwhile, demonstrates a considerable generalizability on a single spectral dataset.

****

Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval

Shupeng Su, Zhisheng Zhong, Chao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3027-3035

Cross-modal hashing encodes the multimedia data into a common binary hash space in which the correlations among the samples from different modalities can be effectively measured. Deep cross-modal hashing further improves the retrieval performance as the deep neural networks can generate more semantic relevant features and hash codes. In this paper, we study the unsupervised deep cross-modal hash coding and propose Deep Joint-Semantics Reconstructing Hashing (DJSRH), which has the following two main advantages. First, to learn binary codes that preserve the neighborhood structure of the original data, DJSRH constructs a novel joint-semantics affinity matrix which elaborately integrates the original neighborhood information from different modalities and accordingly is capable to capture the latent intrinsic semantic affinity for the input multi-modal instances. Second, DJSRH later trains the networks to generate binary codes that maximally reconstruct above joint-semantics relations via the proposed reconstructing framework, which is more competent for the batch-wise training as it reconstructs the specific similarity value unlike the common Laplacian constraint merely preserving the similarity order. Extensive experiments demonstrate the significant improvement by DJSRH in various cross-modal retrieval tasks.

****

Unsupervised Neural Quantization for Compressed-Domain Similarity Search

Stanislav Morozov, Artem Babenko; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2019, pp. 3036-3045

We tackle the problem of unsupervised visual descriptors compression, which is a key ingredient of large-scale image retrieval systems. While the deep learning machinery has benefited literally all computer vision pipelines, the existing state-of-the-art compression methods employ shallow architectures, and we aim to close this gap by our paper. In more detail, we introduce a DNN architecture for the unsupervised compressed-domain retrieval, based on multi-codebook quantization. The proposed architecture is designed to incorporate both fast data encoding and efficient distances computation via lookup tables. We demonstrate the exceptional advantage of our scheme over existing quantization approaches on several datasets of visual descriptors via outperforming the previous state-of-the-art by a large margin.

****

Siamese Networks: The Tale of Two Manifolds

Soumava Kumar Roy, Mehrtash Harandi, Richard Nock, Richard Hartley; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3046-3055

Siamese networks are non-linear deep models that have found their ways into a broad set of problems in learning theory, thanks to their embedding capabilities. In this paper, we study Siamese networks from a new perspective and question the validity of their training procedure. We show that in the majority of cases, the objective of a Siamese network is endowed with an invariance property. Neglecting the invariance property leads to a hindrance in training the Siamese networks. To alleviate this issue, we propose two Riemannian structures and generalize a well-established accelerated stochastic gradient descent method to take into account the proposed Riemannian structures. Our empirical evaluations suggest that by making use of the Riemannian geometry, we achieve state-of-the-art results against several algorithms for the challenging problem of fine-grained image classification.

****

Learning Combinatorial Embedding Networks for Deep Graph Matching

Runzhong Wang, Junchi Yan, Xiaokang Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3056-3065

Graph matching refers to finding node correspondence between graphs, such that the corresponding node and edge's affinity can be maximized. In addition with its NP-completeness nature, another important challenge is effective modeling of the node-wise and structure-wise affinity across graphs and the resulting objective, to guide the matching procedure effectively finding the true matching against noises. To this end, this paper devises an end-to-end differentiable deep network pipeline to learn the affinity for graph matching. It involves a supervised permutation loss regarding with node correspondence to capture the combinatorial nature for graph matching. Meanwhile deep graph embedding models are adopted to parameterize both intra-graph and cross-graph affinity functions, instead of the traditional shallow and simple parametric forms e.g. a Gaussian kernel. The embedding can also effectively capture the higher-order structure beyond second-order edges. The permutation loss model is agnostic to the number of nodes, and the embedding model is shared among nodes such that the network allows for varying numbers of nodes in graphs for training and inference. Moreover, our network is class-agnostic with some generalization capability across different categories. All these features are welcomed for real-world applications. Experiments show its superiority against state-of-the-art graph matching learning methods.

****

Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid

Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, Wayne Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3066-3075

Matching clothing images from customers and online shopping stores has rich applications in E-commerce. Existing algorithms encoded an image as a global feature vector and performed retrieval with the global representation. However, discriminative local information on clothes are submerged in this global representation

, resulting in sub-optimal performance. To address this issue, we propose a novel Graph Reasoning Network (GRNet) on a Similarity Pyramid, which learns similarities between a query and a gallery cloth by using both global and local representations in multiple scales. The similarity pyramid is represented by a Graph of similarity, where nodes represent similarities between clothing components at different scales, and the final matching score is obtained by message passing along edges. In GRNet, graph reasoning is solved by training a graph convolutional network, enabling to align salient clothing components to improve clothing retrieval. To facilitate future researches, we introduce a new benchmark FindFashion, containing rich annotations of bounding boxes, views, occlusions, and cropping. Extensive experiments show that GRNet obtains new state-of-the-art results on two challenging benchmarks, e.g. pushing the top-1, top-20, and top-50 accuracies on DeepFashion to 26%, 64%, and 75% (i.e. 4%, 10%, and 10% absolute improvements), outperforming competitors with large margins. On FindFashion, GRNet achieves considerable improvements on all empirical settings.

****

Wavelet Domain Style Transfer for an Effective Perception-Distortion Tradeoff in Single Image Super-Resolution

Xin Deng, Ren Yang, Mai Xu, Pier Luigi Dragotti; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3076-3085

In single image super-resolution (SISR), given a low-resolution (LR) image, one wishes to find a high-resolution (HR) version of it which is both accurate and photorealistic. Recently, it has been shown that there exists a fundamental tradeoff between low distortion and high perceptual quality, and the generative adversarial network (GAN) is demonstrated to approach the perception-distortion (PD) bound effectively. In this paper, we propose a novel method based on wavelet domain style transfer (WDST), which achieves a better PD tradeoff than the GAN based methods. Specifically, we propose to use 2D stationary wavelet transform (SWT) to decompose one image into low-frequency and high-frequency sub-bands. For the low-frequency sub-band, we improve its objective quality through an enhancement network. For the high-frequency sub-band, we propose to use WDST to effectively improve its perceptual quality. By feat of the perfect reconstruction property of wavelets, these sub-bands can be re-combined to obtain an image which has simultaneously high objective and perceptual quality. The numerical results on various datasets show that our method achieves the best trade-off between the distortion and perceptual quality among the existing state-of-the-art SISR methods.

****

Toward Real-World Single Image Super-Resolution: A New Benchmark and a New Model

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3086-3095

Most of the existing learning-based single image super-resolution (SISR) methods are trained and evaluated on simulated datasets, where the low-resolution (LR) images are generated by applying a simple and uniform degradation (i.e., bicubic downsampling) to their high-resolution (HR) counterparts. However, the degradations in real-world LR images are far more complicated. As a consequence, the SISR models trained on simulated data become less effective when applied to practical scenarios. In this paper, we build a real-world super-resolution (RealSR) dataset where paired LR-HR images on the same scene are captured by adjusting the focal length of a digital camera. An image registration algorithm is developed to progressively align the image pairs at different resolutions. Considering that the degradation kernels are naturally non-uniform in our dataset, we present a Laplacian pyramid based kernel prediction network (LP-KPN), which efficiently learns per-pixel kernels to recover the HR image. Our extensive experiments demonstrate that SISR models trained on our RealSR dataset deliver better visual quality with sharper edges and finer textures on real-world scenes than those trained on simulated datasets. Though our RealSR dataset is built by using only two cameras (Canon 5D3 and Nikon D810), the trained model generalizes well to other camera devices such as Sony a7II and mobile phones.

****

RankSRGAN: Generative Adversarial Networks With Ranker for Image Super-Resolution

Wenlong Zhang, Yihao Liu, Chao Dong, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3096-3105

Generative Adversarial Networks (GAN) have demonstrated the potential to recover realistic details for single image super-resolution (SISR). To further improve the visual quality of super-resolved results, PIRM2018-SR Challenge employed perceptual metrics to assess the perceptual quality, such as PI, NIQE, and Ma. However, existing methods cannot directly optimize these indifferentiable perceptual metrics, which are shown to be highly correlated with human ratings. To address the problem, we propose Super-Resolution Generative Adversarial Networks with Ranker (RankSRGAN) to optimize generator in the direction of perceptual metrics. Specifically, we first train a Ranker which can learn the behavior of perceptual metrics and then introduce a novel rank-content loss to optimize the perceptual quality. The most appealing part is that the proposed method can combine the strengths of different SR methods to generate better results. Extensive experiments show that RankSRGAN achieves visually pleasing results and reaches state-of-the-art performance in perceptual metrics. Project page: https://wenlongzhang0724.github.io/Projects/RankSRGAN

****

Progressive Fusion Video Super-Resolution Network via Exploiting Non-Local Spatio-Temporal Correlations

Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Jiayi Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3106-3115

Most previous fusion strategies either fail to fully utilize temporal information or cost too much time, and how to effectively fuse temporal information from consecutive frames plays an important role in video super-resolution (SR). In this study, we propose a novel progressive fusion network for video SR, which is designed to make better use of spatio-temporal information and is proved to be more efficient and effective than the existing direct fusion, slow fusion or 3D convolution strategies. Under this progressive fusion framework, we further introduce an improved non-local operation to avoid the complex motion estimation and motion compensation (ME&MC) procedures as in previous video SR approaches. Extensive experiments on public datasets demonstrate that our method surpasses state-of-the-art with 0.96 dB in average, and runs about 3 times faster, while requires only about half of the parameters.

****

Deep SR-ITM: Joint Learning of Super-Resolution and Inverse Tone-Mapping for 4K UHD HDR Applications

Soo Ye Kim, Jihyong Oh, Munchurl Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3116-3125

Recent modern displays are now able to render high dynamic range (HDR), high resolution (HR) videos of up to 8K UHD (Ultra High Definition). Consequently, UHD HDR broadcasting and streaming have emerged as high quality premium services. However, due to the lack of original UHD HDR video content, appropriate conversion technologies are urgently needed to transform the legacy low resolution (LR) standard dynamic range (SDR) videos into UHD HDR versions. In this paper, we propose a joint super-resolution (SR) and inverse tone-mapping (ITM) framework, called Deep SR-ITM, which learns the direct mapping from LR SDR video to their HR HDR version. Joint SR and ITM is an intricate task, where high frequency details must be restored for SR, jointly with the local contrast, for ITM. Our network is able to restore fine details by decomposing the input image and focusing on the separate base (low frequency) and detail (high frequency) layers. Moreover, the proposed modulation blocks apply location-variant operations to enhance local contrast. The Deep SR-ITM shows good subjective quality with increased contrast and details, outperforming the previous joint SR-ITM method.

****

Dynamic PET Image Reconstruction Using Nonnegative Matrix Factorization Incorporated With Deep Image Prior

Tatsuya Yokota, Kazuya Kawai, Muneyuki Sakata, Yuichi Kimura, Hidekata Hontani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3126-3135

We propose a method that reconstructs dynamic positron emission tomography (PET) images from given sinograms by using non-negative matrix factorization (NMF) incorporated with a deep image prior (DIP) for appropriately constraining the spatial patterns of resultant images. The proposed method can reconstruct dynamic PET images with higher signal-to-noise ratio (SNR) and blindly decompose an image matrix into pairs of spatial and temporal factors. The former represent homogeneous tissues with different kinetic parameters and the latter represent the time activity curves that are observed in the corresponding homogeneous tissues. We employ U-Nets combined in parallel for DIP and each of the U-nets is used to extract each spatial factor decomposed from the data matrix. Experimental results show that the proposed method outperforms conventional methods and can extract spatial factors that represent the homogeneous tissues.

****

DSIC: Deep Stereo Image Compression

Jerry Liu, Shenlong Wang, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3136-3145

In this paper we tackle the problem of stereo image compression, and leverage the fact that the two images have overlapping fields of view to further compress the representations. Our approach leverages state-of-the-art single-image compression autoencoders and enhances the compression with novel parametric skip functions to feed fully differentiable, disparity-warped features at all levels to the encoder/decoder of the second image. Moreover, we model the probabilistic dependence between the image codes using a conditional entropy model. Our experiments show an impressive 30 - 50% reduction in the second image bitrate at low bitrates compared to deep single-image compression, and a 10 - 20% reduction at higher bitrates.

****

Variable Rate Deep Image Compression With a Conditional Autoencoder

Yoojin Choi, Mostafa El-Khamy, Jungwon Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3146-3154

In this paper, we propose a novel variable-rate learned image compression framework with a conditional autoencoder. Previous learning-based image compression methods mostly require training separate networks for different compression rates so they can yield compressed images of varying quality. In contrast, we train and deploy only one variable-rate image compression network implemented with a conditional autoencoder. We provide two rate control parameters, i.e., the Lagrange multiplier and the quantization bin size, which are given as conditioning variables to the network. Coarse rate adaptation to a target is performed by changing the Lagrange multiplier, while the rate can be further fine-tuned by adjusting the bin size used in quantizing the encoded representation. Our experimental results show that the proposed scheme provides a better rate-distortion trade-off than the traditional variable-rate image compression codecs such as JPEG2000 and BPG. Our model also shows comparable and sometimes better performance than the state-of-the-art learned image compression models that deploy multiple networks trained for varying rates.

****

Real Image Denoising With Feature Attention

Saeed Anwar, Nick Barnes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3155-3164

Deep convolutional neural networks perform better on images containing spatially invariant noise (synthetic noise); however, its performance is limited on real-noisy photographs and requires multiple stage network modeling. To advance the practicability of the denoising algorithms, this paper proposes a novel single-stage blind real image denoising network (RIDNet) by employing a modular architecture. We use residual on the residual structure to ease the flow of low-frequency information and apply feature attention to exploit the channel dependencies. Furthermore, the evaluation in terms of quantitative metrics and visual quality on

three synthetic and four real noisy datasets against 19 state-of-the-art algorithms demonstrate the superiority of our RIDNet.
****

Noise Flow: Noise Modeling With Conditional Normalizing Flows
Abdelrahman Abdelhamed, Marcus A. Brubaker, Michael S. Brown; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3165-3173

Modeling and synthesizing image noise is an important aspect in many computer vision applications. The long-standing additive white Gaussian and heteroscedastic (signal-dependent) noise models widely used in the literature provide only a coarse approximation of real sensor noise. This paper introduces Noise Flow, a powerful and accurate noise model based on recent normalizing flow architectures. Noise Flow combines well-established basic parametric noise models (e.g., signal-dependent noise) with the flexibility and expressiveness of normalizing flow networks. The result is a single, comprehensive, compact noise model containing fewer than 2500 parameters yet able to represent multiple cameras and gain factors. Noise Flow dramatically outperforms existing noise models, with 0.42 nats/pixel improvement over the camera-calibrated noise level functions, which translates to 52% improvement in the likelihood of sampled noise. Noise Flow represents the first serious attempt to go beyond simple parametric models to one that leverages the power of deep learning and data-driven noise distributions.
****

Bottleneck Potentials in Markov Random Fields
Ahmed Abbas, Paul Swoboda; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3175-3184

We consider general discrete Markov Random Fields(MRFs) with additional bottleneck potentials which penalize the maximum (instead of the sum) over local potential value taken by the MRF-assignment. Bottleneck potentials or analogous constructions have been considered in (i) combinatorial optimization (e.g. bottleneck shortest path problem, the minimum bottleneck spanning tree problem, bottleneck function minimization in greedoids), (ii) inverse problems with L_ infinity -norm regularization and (iii) valued constraint satisfaction on the (min,max)-pre-semirings. Bottleneck potentials for general discrete MRFs are a natural generalization of the above direction of modeling work to Maximum-A-Posteriori (MAP) inference in MRFs. To this end we propose MRFs whose objective consists of two parts: terms that factorize according to (i) (min,+), i.e. potentials as in plain MRFs, and (ii) (min,max), i.e. bottleneck potentials. To solve the ensuing inference problem, we propose high-quality relaxations and efficient algorithms for solving them. We empirically show efficacy of our approach on large scale seismic horizon tracking problems.
****

Seeing Motion in the Dark
Chen Chen, Qifeng Chen, Minh N. Do, Vladlen Koltun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3185-3194

Deep learning has recently been applied with impressive results to extreme low-light imaging. Despite the success of single-image processing, extreme low-light video processing is still intractable due to the difficulty of collecting raw video data with corresponding ground truth. Collecting long-exposure ground truth, as was done for single-image processing, is not feasible for dynamic scenes. In this paper, we present deep processing of very dark raw videos: on the order of one lux of illuminance. To support this line of work, we collect a new dataset of raw low-light videos, in which high-resolution raw data is captured at video rate. At this level of darkness, the signal-to-noise ratio is extremely low (negative if measured in dB) and the traditional image processing pipeline generally breaks down. A new method is presented to address this challenging problem. By carefully designing a learning-based pipeline and introducing a new loss function to encourage temporal stability, we train a siamese network on static raw videos, for which ground truth is available, such that the network generalizes to videos of dynamic scenes at test time. Experimental results demonstrate that the presented approach outperforms state-of-the-art models for burst processing, per-frame processing, and blind temporal consistency.

****

SENSE: A Shared Encoder Network for Scene-Flow Estimation
Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3195-3204

We introduce a compact network for holistic scene flow estimation, called SENSE, which shares common encoder features among four closely-related tasks: optical flow estimation, disparity estimation from stereo, occlusion estimation, and semantic segmentation. Our key insight is that sharing features makes the network more compact, induces better feature representations, and can better exploit interactions among these tasks to handle partially labeled data. With a shared encoder, we can flexibly add decoders for different tasks during training. This modular design leads to a compact and efficient model at inference time. Exploiting the interactions among these tasks allows us to introduce distillation and self-supervised losses in addition to supervised losses, which can better handle partially labeled real-world data. SENSE achieves state-of-the-art results on several optical flow benchmarks and runs as fast as networks specifically designed for optical flow. It also compares favorably against the state of the art on stereo and scene flow, while consuming much less memory.

****

Adversarial Feedback Loop
Firas Shama, Roey Mechrez, Alon Shoshan, Lihi Zelnik-Manor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3205-3214

Thanks to their remarkable generative capabilities, GANs have gained great popularity, and are used abundantly in state-of-the-art methods and applications. In a GAN based model, a discriminator is trained to learn the real data distribution. To date, it has been used only for training purposes, where it's utilized to train the generator to provide real-looking outputs. In this paper we propose a novel method that makes an explicit use of the discriminator in test-time, in a feedback manner in order to improve the generator results. To the best of our knowledge it is the first time a discriminator is involved in test-time. We claim that the discriminator holds significant information on the real data distribution, that could be useful for test-time as well, a potential that has not been explored before. The approach we propose does not alter the conventional training stage. At test-time, however, it transfers the output from the generator into the discriminator, and uses feedback modules (convolutional blocks) to translate the features of the discriminator layers into corrections to the features of the generator layers, which are used eventually to get a better generator result. Our method can contribute to both conditional and unconditional GANs. As demonstrated by our experiments, it can improve the results of state-of-the-art networks for super-resolution, and image generation.

****

Dynamic-Net: Tuning the Objective Without Re-Training for Synthesis Tasks
Alon Shoshan, Roey Mechrez, Lihi Zelnik-Manor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3215-3223

One of the key ingredients for successful optimization of modern CNNs is identifying a suitable objective. To date, the objective is fixed a-priori at training time, and any variation to it requires re-training a new network. In this paper we present a first attempt at alleviating the need for re-training. Rather than fixing the network at training time, we train a "Dynamic-Net" that can be modified at inference time. Our approach considers an "objective-space" as the space of all linear combinations of two objectives, and the Dynamic-Net is emulating the traversing of this objective-space at test-time, without any further training. We show that this upgrades pre-trained networks by providing an out-of-learning extension, while maintaining the performance quality. The solution we propose is fast and allows a user to interactively modify the network, in real-time, in order to obtain the result he/she desires. We show the benefits of such an approach via several different applications.

****

AutoGAN: Neural Architecture Search for Generative Adversarial Networks

Xinyu Gong, Shiyu Chang, Yifan Jiang, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3224-3234

Neural architecture search (NAS) has witnessed prevailing success in image classification and (very recently) segmentation tasks. In this paper, we present the first preliminary study on introducing the NAS algorithm to generative adversarial networks (GANs), dubbed AutoGAN. The marriage of NAS and GANs faces its unique challenges. We define the search space for the generator architectural variations and use an RNN controller to guide the search, with parameter sharing and dynamic-resetting to accelerate the process. Inception score is adopted as the reward, and a multi-level search strategy is introduced to perform NAS in a progressive way. Experiments validate the effectiveness of AutoGAN on the task of unconditional image generation. Specifically, our discovered architectures achieve highly competitive performance compared to current state-of-the-art hand-crafted GANs, e.g., setting new state-of-the-art FID scores of 12.42 on CIFAR-10, and 31.01 on STL-10, respectively. We also conclude with a discussion of the current limitations and future potential of AutoGAN. The code is available at https://github.com/TAMU-VITA/AutoGAN

****

Co-Evolutionary Compression for Unpaired Image Translation

Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen, Chunjing Xu, Qi Tian, Chang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3235-3244

Generative adversarial networks (GANs) have been successfully used for considerable computer vision tasks, especially the image-to-image translation. However, generators in these networks are of complicated architectures with large number of parameters and huge computational complexities. Existing methods are mainly designed for compressing and speeding-up deep neural networks in the classification task, and cannot be directly applied on GANs for image translation, due to their different objectives and training procedures. To this end, we develop a novel co-evolutionary approach for reducing their memory usage and FLOPs simultaneously. In practice, generators for two image domains are encoded as two populations and synergistically optimized for investigating the most important convolution filters iteratively. Fitness of each individual is calculated using the number of parameters, a discriminator-aware regularization, and the cycle consistency. Extensive experiments conducted on benchmark datasets demonstrate the effectiveness of the proposed method for obtaining compact and effective generators.

****

Self-Supervised Representation Learning From Multi-Domain Data

Zeyu Feng, Chang Xu, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3245-3255

We present an information-theoretically motivated constraint for self-supervised representation learning from multiple related domains. In contrast to previous self-supervised learning methods, our approach learns from multiple domains, which has the benefit of decreasing the build-in bias of individual domain, as well as leveraging information and allowing knowledge transfer across multiple domains. The proposed mutual information constraints encourage neural network to extract common invariant information across domains and to preserve peculiar information of each domain simultaneously. We adopt tractable upper and lower bounds of mutual information to make the proposed constraints solvable. The learned representation is more unbiased and robust toward the input images. Extensive experimental results on both multi-domain and large-scale datasets demonstrate the necessity and advantage of multi-domain self-supervised learning with mutual information constraints. Representations learned in our framework on state-of-the-art methods achieve improved performance than those learned on a single domain.

****

Controlling Neural Networks via Energy Dissipation

Michael Moeller, Thomas Mollenhoff, Daniel Cremers; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3256-3265

The last decade has shown a tremendous success in solving various computer vision problems with the help of deep learning techniques. Lately, many works have demonstrated that learning-based approaches with suitable network architectures even exhibit superior performance for the solution of (ill-posed) image reconstruction problems such as deblurring, super-resolution, or medical image reconstruction. The drawback of purely learning-based methods, however, is that they cannot provide provable guarantees for the trained network to follow a given data formation process during inference. In this work we propose energy dissipating networks that iteratively compute a descent direction with respect to a given cost function or energy at the currently estimated reconstruction. Therefore, an adaptive step size rule such as a line-search, along with a suitable number of iterations can guarantee the reconstruction to follow a given data formation model encoded in the energy to arbitrary precision, and hence control the model's behavior even during test time. We prove that under standard assumptions, descent using the direction predicted by the network converges (linearly) to the global minimum of the energy. We illustrate the effectiveness of the proposed approach in experiments on single image super resolution and computed tomography (CT) reconstruction, and further illustrate extensions to convex feasibility problems.

****

Indices Matter: Learning to Index for Deep Image Matting

Hao Lu, Yutong Dai, Chunhua Shen, Songcen Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3266-3275

We show that existing upsampling operators can be unified using the notion of the index function. This notion is inspired by an observation in the decoding process of deep image matting where indices-guided unpooling can often recover boundary details considerably better than other upsampling operators such as bilinear interpolation. By viewing the indices as a function of the feature map, we introduce the concept of 'learning to index', and present a novel index-guided encoder-decoder framework where indices are self-learned adaptively from data and are used to guide the pooling and upsampling operators, without extra training supervision. At the core of this framework is a flexible network module, termed IndexNet, which dynamically generates indices conditioned on the feature map. Due to its flexibility, IndexNet can be used as a plug-in applying to almost all off-the-shelf convolutional networks that have coupled downsampling and upsampling stages. We demonstrate the effectiveness of IndexNet on the task of natural image matting where the quality of learned indices can be visually observed from predicted alpha mattes. Results on the Composition-1k matting dataset show that our model built on MobileNetv2 exhibits at least 16.1% improvement over the seminal VGG-16 based deep matting baseline, with less training data and lower model capacity. Code and models have been made available at: https://tinyurl.com/IndexNetV1.

****

LAP-Net: Level-Aware Progressive Network for Image Dehazing

Yunan Li, Qiguang Miao, Wanli Ouyang, Zhenxin Ma, Huijuan Fang, Chao Dong, Yining Quan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3276-3285

In this paper, we propose a level-aware progressive network (LAP-Net) for single image dehazing. Unlike previous multi-stage algorithms that generally learn in a coarse-to-fine fashion, each stage of LAP-Net learns different levels of haze with different supervision. Then the network can progressively learn the gradually aggravating haze. With this design, each stage can focus on a region with specific haze level and restore clear details. To effectively fuse the results of varying haze levels at different stages, we develop an adaptive integration strategy to yield the final dehazed image. This strategy is achieved by a hierarchical integration scheme, which is in cooperation with the memory network and the domain knowledge of dehazing to highlight the best-restored regions of each stage. Extensive experiments on both real-world images and two dehazing benchmarks validate the effectiveness of our proposed method.

****

Attention Augmented Convolutional Networks

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, Quoc V. Le; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3286-3295

Convolutional networks have enjoyed much success in many computer vision applications. The convolution operation however has a significant weakness in that it only operates on a local neighbourhood, thus missing global information. Self-attention, on the other hand, has emerged as a recent advance to capture long range interactions, but has mostly been applied to sequence modeling and generative modeling tasks. In this paper, we propose to augment convolutional networks with self-attention by concatenating convolutional feature maps with a set of feature maps produced via a novel relative self-attention mechanism. In particular, we extend previous work on relative self-attention over sequences to images and discuss a memory efficient implementation. Unlike Squeeze-and-Excitation, which performs attention over the channels and ignores spatial information, our self-attention mechanism attends jointly to both features and spatial locations while preserving translation equivariance. We find that Attention Augmentation leads to consistent improvements in image classification on ImageNet and object detection on COCO across many different models and scales, including ResNets and a state-of-the art mobile constrained network, while keeping the number of parameters similar. In particular, our method achieves a 1.3% top-1 accuracy improvement on ImageNet classification over a ResNet50 baseline and outperforms other attention mechanisms for images such as Squeeze-and-Excitation. It also achieves an improvement of 1.4 AP in COCO Object Detection on top of a RetinaNet baseline.

****

MetaPruning: Meta Learning for Automatic Neural Network Channel Pruning

Zechun Liu, Haoyuan Mu, Xiangyu Zhang, Zichao Guo, Xin Yang, Kwang-Ting Cheng, Jian Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3296-3305

In this paper, we propose a novel meta learning approach for automatic channel pruning of very deep neural networks. We first train a PruningNet, a kind of meta network, which is able to generate weight parameters for any pruned structure given the target network. We use a simple stochastic structure sampling method for training the PruningNet. Then, we apply an evolutionary procedure to search for good-performing pruned networks. The search is highly efficient because the weights are directly generated by the trained PruningNet and we do not need any finetuning at search time. With a single PruningNet trained for the target network, we can search for various Pruned Networks under different constraints with little human participation. Compared to the state-of-the-art pruning methods, we have demonstrated superior performances on MobileNet V1/V2 and ResNet. Codes are available on https://github.com/liuzechun/MetaPruning.

****

Accelerate CNN via Recursive Bayesian Pruning

Yuefu Zhou, Ya Zhang, Yanfeng Wang, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3306-3315

Channel Pruning, widely used for accelerating Convolutional Neural Networks, is an NP-hard problem due to the inter-layer dependency of channel redundancy. Existing methods generally ignored the above dependency for computation simplicity. To solve the problem, under the Bayesian framework, we here propose a layer-wise Recursive Bayesian Pruning method (RBP). A new dropout-based measurement of redundancy, which facilitate the computation of posterior assuming inter-layer dependency, is introduced. Specifically, we model the noise across layers as a Markov chain and target its posterior to reflect the inter-layer dependency. Considering the closed form solution for posterior is intractable, we derive a sparsity-inducing Dirac-like prior which regularizes the distribution of the designed noise to automatically approximate the posterior. Compared with the existing methods, no additional overhead is required when the inter-layer dependency assumed. The redundant channels can be simply identified by tiny dropout noise and directly pruned layer by layer. Experiments on popular CNN architectures have shown that the proposed method outperforms several state-of-the-arts. Particularly, we achieve up to 5.0x, 2.2x and 1.7x FLOPs reduction with little accuracy loss on the large scale dataset ILSVRC2012 for VGG16, ResNet50 and MobileNetV2, respectively.

****

HBONet: Harmonious Bottleneck on Two Orthogonal Dimensions

Duo Li, Aojun Zhou, Anbang Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3316-3325

MobileNets, a class of top-performing convolutional neural network architectures in terms of accuracy and efficiency trade-off, are increasingly used in many resource-aware vision applications. In this paper, we present Harmonious Bottleneck on two Orthogonal dimensions (HBO), a novel architecture unit, specially tailored to boost the accuracy of extremely lightweight MobileNets at the level of less than 40 MFLOPs. Unlike existing bottleneck designs that mainly focus on exploring the interdependencies among the channels of either groupwise or depthwise convolutional features, our HBO improves bottleneck representation while maintaining similar complexity via jointly encoding the feature interdependencies across both spatial and channel dimensions. It has two reciprocal components, namely spatial contraction-expansion and channel expansion-contraction, nested in a bilaterally symmetric structure. The combination of two interdependent transformations performing on orthogonal dimensions of feature maps enhances the representation and generalization ability of our proposed module, guaranteeing compelling performance with limited computational resource and power. By replacing the original bottlenecks in MobileNetV2 backbone with HBO modules, we construct HBONets which are evaluated on ImageNet classification, PASCAL VOC object detection and Market-1501 person re-identification. Extensive experiments show that with the severe constraint of computational budget our models outperform MobileNetV2 counterparts by remarkable margins of at most 6.6%, 6.3% and 5.0% on the above benchmarks respectively. Code and pretrained models are available at https://github.com/d-li14/HBONet.

****

O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks

Jinchi Huang, Lie Qu, Rongfei Jia, Binqiang Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3326-3334

This paper proposes a novel noisy label detection approach, named O2U-net, for deep neural networks without human annotations. Different from prior work which requires specifically designed noise-robust loss functions or networks, O2U-net is easy to implement but effective. It only requires adjusting the hyper-parameters of the deep network to make its status transfer from overfitting to underfitting (O2U) cyclically. The losses of each sample are recorded during iterations. The higher the normalized average loss of a sample, the higher the probability of being noisy labels. O2U-net is naturally compatible with active learning and other human annotation approaches. This introduces extra flexibility for learning with noisy labels. We conduct sufficient experiments on multiple datasets in various settings. The experimental results prove the state-of-the-art of O2S-net.

****

Continual Learning by Asymmetric Loss Approximation With Single-Side Overestimation

Dongmin Park, Seokil Hong, Bohyung Han, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3335-3344

Catastrophic forgetting is a critical challenge in training deep neural networks. Although continual learning has been investigated as a countermeasure to the problem, it often suffers from the requirements of additional network components and the limited scalability to a large number of tasks. We propose a novel approach to continual learning by approximating a true loss function using an asymmetric quadratic function with one of its sides overestimated. Our algorithm is motivated by the empirical observation that the network parameter updates affect the target loss functions asymmetrically. In the proposed continual learning framework, we estimate an asymmetric loss function for the tasks considered in the past through a proper overestimation of its unobserved sides in training new tasks, while deriving the accurate model parameter for the observable sides. In contrast to existing approaches, our method is free from the side effects and achieve s the state-of-the-art accuracy that is even close to the upper-bound performance on several challenging benchmark datasets.

****

Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation

Weifeng Ge, Sheng Guo, Weilin Huang, Matthew R. Scott; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3345-3354

Weakly-supervised instance segmentation aims to detect and segment object instances precisely, given image-level labels only. Unlike previous methods which are composed of multiple offline stages, we propose Sequential Label Propagation and Enhancement Networks (referred as Label-PEnet) that progressively transforms image-level labels to pixel-wise labels in a coarse-to-fine manner. We design four cascaded modules including multi-label classification, object detection, instance refinement and instance segmentation, which are implemented sequentially by sharing the same backbone. The cascaded pipeline is trained alternatively with a curriculum learning strategy that generalizes labels from high level images to low-level pixels gradually with increasing accuracy. In addition, we design a proposal calibration module to explore the ability of classification networks to find key pixels that identify object parts, which serves as a post validation strategy running in the inverse order. We evaluate the efficiency of our Label-PEnet in mining instance masks on standard benchmarks: PASCAL VOC 2007 and 2012. Experimental results show that Label-PEnet outperforms the state-of-art algorithms by a clear margin, and obtains comparable performance even with fully supervised approaches.

****

LIP: Local Importance-Based Pooling

Ziteng Gao, Limin Wang, Gangshan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3355-3364

Spatial downsampling layers are favored in convolutional neural networks (CNNs) to downscale feature maps for larger receptive fields and less memory consumption. However, for discriminative tasks, there is a possibility that these layers lose the discriminative details due to improper pooling strategies, which could hinder the learning process and eventually result in suboptimal models. In this paper, we present a unified framework over the existing downsampling layers (e.g., average pooling, max pooling, and strided convolution) from a local importance view. In this framework, we analyze the issues of these widely-used pooling layers and figure out the criteria for designing an effective downsampling layer. According to this analysis, we propose a conceptually simple, general, and effective pooling layer based on local importance modeling, termed as Local Importance-based Pooling (LIP). LIP can automatically enhance discriminative features during the downsampling procedure by learning adaptive importance weights based on inputs. Experiment results show that LIP consistently yields notable gains with different depths and different architectures on ImageNet classification. In the challenging MS COCO dataset, detectors with our LIP-ResNets as backbones obtain a consistent improvement (>=1.4%) over the vanilla ResNets, and especially achieve the current state-of-the-art performance in detecting small objects under the single-scale testing scheme.

****

Global Feature Guided Local Pooling

Takumi Kobayashi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3365-3374

In deep convolutional neural networks (CNNs), local pooling operation is a key building block to effectively downsize feature maps for reducing computation cost as well as increasing robustness against input variation. There are several types of pooling operation, such as average/max-pooling, from which one has to be manually selected for building CNNs. The optimal pooling type would be dependent on characteristics of features in CNNs and classification tasks, making it hard to find out the proper pooling module in advance. In this paper, we propose a flexible pooling method which adaptively tunes the pooling functionality based on input features without manually fixing it beforehand. In the proposed method, th e parameterized pooling form is derived from a probabilistic perspective to flexibly represent various types of pooling and then the parameters are estimated by means of global statistics in the input feature map. Thus, the proposed local pooling guided by global features effectively works in the CNNs trained in an end-to-end manner. The experimental results on image classification tasks demonstrate the effectiveness of the proposed pooling method in various deep CNNs.

****

Conditional Coupled Generative Adversarial Networks for Zero-Shot Domain Adaptation

Jinghua Wang, Jianmin Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3375-3384

Machine learning models trained in one domain perform poorly in the other domains due to the existence of domain shift. Domain adaptation techniques solve this problem by training transferable models from the label-rich source domain to the label-scarce target domain. Unfortunately, a majority of the existing domain adaptation techniques rely on the availability of the target-domain data, and thus limit their applications to a small community across few computer vision problems. In this paper, we tackle the challenging zero-shot domain adaptation (ZSDA) problem, where the target-domain data is non-available in the training stage. For this purpose, we propose conditional coupled generative adversarial networks (CoCoGAN) by extending the coupled generative adversarial networks (CoGAN) into a conditioning model. Compared with the existing state of the arts, our proposed CoCoGAN is able to capture the joint distribution of dual-domain samples in two different tasks, i.e. the relevant task (RT) and an irrelevant task (IRT). We train the CoCoGAN with both source-domain samples in RT and the dual-domain samples in IRT to complete the domain adaptation. While the former provide the high-level concepts of the non-available target-domain data, the latter carry the sharing correlation between the two domains in RT and IRT. To train the CoCoGAN in the absence of the target-domain data for RT, we propose a new supervisory signal, i.e. the alignment between representations across tasks. Extensive experiments carried out demonstrate that our proposed CoCoGAN outperforms existing state of the arts in image classifications.

****

Adversarial Defense by Restricting the Hidden Space of Deep Neural Networks

Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3385-3394

Deep neural networks are vulnerable to adversarial attacks which can fool them by adding minuscule perturbations to the input images. The robustness of existing defenses suffers greatly under white-box attack settings, where an adversary has full knowledge about the network and can iterate several times to find strong perturbations. We observe that the main reason for the existence of such perturbations is the close proximity of different class samples in the learned feature space. This allows model decisions to be totally changed by adding an imperceptible perturbation in the inputs. To counter this, we propose to class-wise disentangle the intermediate feature representations of deep networks. Specifically, we force the features for each class to lie inside a convex polytope that is maximally separated from the polytopes of other classes. In this manner, the network is forced to learn distinct and distant decision regions for each class. We observe that this simple constraint on the features greatly enhances the robustness of learned models, even against the strongest white-box attacks, without degrading the classification performance on clean images. We report extensive evaluations in both black-box and white-box attack scenarios and show significant gains in comparison to state-of-the art defenses.

****

Hyperpixel Flow: Semantic Correspondence With Multi-Layer Neural Features

Juhong Min, Jongmin Lee, Jean Ponce, Minsu Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3395-3404

Establishing visual correspondences under large intra-class variations requires analyzing images at different levels, from features linked to semantics and cont ext to local patterns, while being invariant to instance-specific details. To tackle these challenges, we represent images by "hyperpixels" that leverage a small number of relevant features selected among early to late layers of a convolutional neural network. Taking advantage of the condensed features of hyperpixels, we develop an effective real-time matching algorithm based on Hough geometric voting. The proposed method, hyperpixel flow, sets a new state of the art on three standard benchmarks as well as a new dataset, SPair-71k, which contains a significantly larger number of image pairs than existing datasets, with more accurate and richer annotations for in-depth analysis.

****

Information Entropy Based Feature Pooling for Convolutional Neural Networks

Weitao Wan, Jiansheng Chen, Tianpeng Li, Yiqing Huang, Jingqi Tian, Cheng Yu, Youze Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3405-3414

In convolutional neural networks (CNNs), we propose to estimate the importance of a feature vector at a spatial location in the feature maps by the network's uncertainty on its class prediction, which can be quantified using the information entropy. Based on this idea, we propose the entropy-based feature weighting method for semantics-aware feature pooling which can be readily integrated into various CNN architectures for both training and inference. We demonstrate that such a location-adaptive feature weighting mechanism helps the network to concentrate on semantically important image regions, leading to improvements in the large-scale classification and weakly-supervised semantic segmentation tasks. Furthermore, the generated feature weights can be utilized in visual tasks such as weakly-supervised object localization. We conduct extensive experiments on different datasets and CNN architectures, outperforming recently proposed pooling methods and attention mechanisms in ImageNet classification as well as achieving state-of-the-arts in weakly-supervised semantic segmentation on PASCAL VOC 2012 dataset.

****

Patchwork: A Patch-Wise Attention Network for Efficient Object Detection and Segmentation in Video Streams

Yuning Chai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3415-3424

Recent advances in single-frame object detection and segmentation techniques have motivated a wide range of works to extend these methods to process video streams. In this paper, we explore the idea of hard attention aimed for latency-sensitive applications. Instead of reasoning about every frame separately, our method selects and only processes a small sub-window of the frame. Our technique then makes predictions for the full frame based on the sub-windows from previous frames and the update from the current sub-window. The latency reduction by this hard attention mechanism comes at the cost of degraded accuracy. We made two contributions to address this. First, we propose a specialized memory cell that recovers lost context when processing sub-windows. Secondly, we adopt a Q-learning-based policy training strategy that enables our approach to intelligently select the sub-windows such that the staleness in the memory hurts the performance the least. Our experiments suggest that our approach reduces the latency by approximately four times without significantly sacrificing the accuracy on the ImageNet VID video object detection dataset and the DAVIS video object segmentation dataset. We further demonstrate that we can reinvest the saved computation into other parts of the network, and thus resulting in an accuracy increase at a comparable computational cost as the original system and beating other recently proposed state-of-the-art methods in the low latency range.

****

AttentionRNN: A Structured Spatial Attention Mechanism

Siddhesh Khandelwal, Leonid Sigal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3425-3434

Visual attention mechanisms have proven to be integrally important constituent components of many modern deep neural architectures. They provide an efficient and effective way to utilize visual information selectively, which has shown to be especially valuable in multi-modal learning tasks. However, all prior attention frameworks lack the ability to explicitly model structural dependencies among attention variables, making it difficult to predict consistent attention masks. In this paper we develop a novel structured spatial attention mechanism which is end-to-end trainable and can be integrated with any feed-forward convolutional neural network. This proposed AttentionRNN layer explicitly enforces structure over the spatial attention variables by sequentially predicting attention values in the spatial mask in a bi-directional raster-scan and inverse raster-scan order. As a result, each attention value depends not only on local image or contextual information, but also on the previously predicted attention values. Our experiments show consistent quantitative and qualitative improvements on a variety of recognition tasks and datasets; including image categorization, question answering and image generation.

****

Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks With Octave Convolution
Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3435-3444

In natural images, information is conveyed at different frequencies where higher frequencies are usually encoded with fine details and lower frequencies are usually encoded with global structures. Similarly, the output feature maps of a convolution layer can also be seen as a mixture of information at different frequencies. In this work, we propose to factorize the mixed feature maps by their frequencies, and design a novel Octave Convolution (OctConv) operation to store and process feature maps that vary spatially "slower" at a lower spatial resolution reducing both memory and computation cost. Unlike existing multi-scale methods, OctConv is formulated as a single, generic, plug-and-play convolutional unit that can be used as a direct replacement of (vanilla) convolutions without any adjustments in the network architecture. It is also orthogonal and complementary to methods that suggest better topologies or reduce channel-wise redundancy like group or depth-wise convolutions. We experimentally show that by simply replacing convolutions with OctConv, we can consistently boost accuracy for both image and video recognition tasks, while reducing memory and computational cost. An OctConv-equipped ResNet-152 can achieve 82.9% top-1 classification accuracy on ImageNet with merely 22.2 GFLOPs.

****

Domain Intersection and Domain Difference
Sagie Benaim, Michael Khaitov, Tomer Galanti, Lior Wolf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3445-3453

We present a method for recovering the shared content between two visual domains as well as the content that is unique to each domain. This allows us to map from one domain to the other, in a way in which the content that is specific for the first domain is removed and the content that is specific for the second is imported from any image in the second domain. In addition, our method enables generation of images from the intersection of the two domains as well as their union, despite having no such samples during training. The method is shown analytically to contain all the sufficient and necessary constraints. It also outperforms the literature methods in an extensive set of experiments.

****

Learned Video Compression
Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, Lubomir Bourdev; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3454-3463

We present a new algorithm for video coding, learned end-to-end for the low-latency mode. In this setting, our approach outperforms all existing video codecs across nearly the entire bitrate range. To our knowledge, this is the first ML-based method to do so. We evaluate our approach on standard video compression test sets of varying resolutions, and benchmark against all mainstream commercial codecs in the low-latency mode. On standard-definition videos, HEVC/H.265, AVC/H.26

4 and VP9 typically produce codes up to 60% larger than our algorithm. On high-definition 1080p videos, H.265 and VP9 typically produce codes up to 20% larger, and H.264 up to 35% larger. Furthermore, our approach does not suffer from blocking artifacts and pixelation, and thus produces videos that are more visually pleasing. We propose two main contributions. The first is a novel architecture for video compression, which (1) generalizes motion estimation to perform any learned compensation beyond simple translations, (2) rather than strictly relying on previously transmitted reference frames, maintains a state of arbitrary information learned by the model, and (3) enables jointly compressing all transmitted signals (such as optical flow and residual). Secondly, we present a framework for ML-based spatial rate control --- a mechanism for assigning variable bitrates across space for each frame. This is a critical component for video coding, which to our knowledge had not been developed within a machine learning setting.
****

Local Relation Networks for Image Recognition
Han Hu, Zheng Zhang, Zhenda Xie, Stephen Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3464-3473
The convolution layer has been the dominant feature extractor in computer vision for years. However, the spatial aggregation in convolution is basically a pattern matching process that applies fixed filters which are inefficient at modeling visual elements with varying spatial distributions. This paper presents a new image feature extractor, called the local relation layer, that adaptively determines aggregation weights based on the compositional relationship of local pixel pairs. With this relational approach, it can composite visual elements into higher-level entities in a more efficient manner that benefits semantic inference. A network built with local relation layers, called the Local Relation Network (LR-Net), is found to provide greater modeling capacity than its counterpart built with regular convolution on large-scale recognition tasks such as ImageNet classification.
****

DiscoNet: Shapes Learning on Disconnected Manifolds for 3D Editing
Eloi Mehr, Ariane Jourdan, Nicolas Thome, Matthieu Cord, Vincent Guitteny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3474-3483
Editing 3D models is a very challenging task, as it requires complex interactions with the 3D shape to reach the targeted design, while preserving the global consistency and plausibility of the shape. In this work, we present an intelligent and user-friendly 3D editing tool, where the edited model is constrained to lie onto a learned manifold of realistic shapes. Due to the topological variability of real 3D models, they often lie close to a disconnected manifold, which cannot be learned with a common learning algorithm. Therefore, our tool is based on a new deep learning model, DiscoNet, which extends 3D surface autoencoders in two ways. Firstly, our deep learning model uses several autoencoders to automatically learn each connected component of a disconnected manifold, without any supervision. Secondly, each autoencoder infers the output 3D surface by deforming a pre-learned 3D template specific to each connected component. Both advances translate into improved 3D synthesis, thus enhancing the quality of our 3D editing tool.
****

Deep Residual Learning in the JPEG Transform Domain
Max Ehrlich, Larry S. Davis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3484-3493
We introduce a general method of performing Residual Network inference and learning in the JPEG transform domain that allows the network to consume compressed images as input. Our formulation leverages the linearity of the JPEG transform to redefine convolution and batch normalization with a tune-able numerical approximation for ReLu. The result is mathematically equivalent to the spatial domain network up to the ReLu approximation accuracy. A formulation for image classification and a model conversion algorithm for spatial domain networks are given as examples of the method. We show that the sparsity of the JPEG format allows for f

aster processing of images with little to no penalty in the network accuracy.

****

Approximated Bilinear Modules for Temporal Modeling

Xinqi Zhu, Chang Xu, Langwen Hui, Cewu Lu, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3494-3503

We consider two less-emphasized temporal properties of video: 1. Temporal cues are fine-grained; 2. Temporal modeling needs reasoning. To tackle both problems at once, we exploit approximated bilinear modules (ABMs) for temporal modeling. There are two main points making the modules effective: two-layer MLPs can be seen as a constraint approximation of bilinear operations, thus can be used to construct deep ABMs in existing CNNs while reusing pretrained parameters; frame features can be divided into static and dynamic parts because of visual repetition in adjacent frames, which enables temporal modeling to be more efficient. Multiple ABM variants and implementations are investigated, from high performance to high efficiency. Specifically, we show how two-layer subnets in CNNs can be converted to temporal bilinear modules by adding an auxiliary-branch. Besides, we introduce snippet sampling and shifting inference to boost sparse-frame video classification performance. Extensive ablation studies are conducted to show the effectiveness of proposed techniques. Our models can outperform most state-of-the-art methods on Something-Something v1 and v2 datasets without Kinetics pretraining, and are also competitive on other YouTube-like action recognition datasets. Our code is available on https://github.com/zhuxinqimac/abm-pytorch.

****

Customizing Student Networks From Heterogeneous Teachers via Adaptive Knowledge Amalgamation

Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, Mingli Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3504-3513

A massive number of well-trained deep networks have been released by developers online. These networks may focus on different tasks and in many cases are optimized for different datasets. In this paper, we study how to exploit such heterogeneous pre-trained networks, known as teachers, so as to train a customized student network that tackles a set of selective tasks defined by the user. We assume no human annotations are available, and each teacher may be either single- or multi-task. To this end, we introduce a dual-step strategy that first extracts the task-specific knowledge from the heterogeneous teachers sharing the same sub-task, and then amalgamates the extracted knowledge to build the student network. To facilitate the training, we employ a selective learning scheme where, for each unlabelled sample, the student learns adaptively from only the teacher with the least prediction ambiguity. We evaluate the proposed approach on several datasets and the experimental results demonstrate that the student, learned by such adaptive knowledge amalgamation, achieves performances even better than those of the teachers.

****

Data-Free Learning of Student Networks

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3514-3522

Learning portable neural networks is very essential for computer vision for the purpose that pre-trained heavy deep models can be well applied on edge devices such as mobile phones and micro sensors. Most existing deep neural network compression and speed-up methods are very effective for training compact deep models, when we can directly access the training dataset. However, training data for the given deep network are often unavailable due to some practice problems (e.g. privacy, legal issue, and transmission), and the architecture of the given network are also unknown except some interfaces. To this end, we propose a novel framework for training efficient deep neural networks by exploiting generative adversarial networks (GANs). To be specific, the pre-trained teacher networks are regarded as a fixed discriminator and the generator is utilized for derivating training samples which can obtain the maximum response on the discriminator. Then, an efficient network with smaller model size and computational complexity is trained using the generated data and the teacher network, simultaneously. Efficient student networks learned using the proposed Data-Free Learning (DFL) method achieve 92.22% and 74.47% accuracies without any training data on the CIFAR-10 and CIFAR-100 datasets, respectively. Meanwhile, our student network obtains an 80.56% accuracy on the CelebA benchmark.

****

Deep Closest Point: Learning Representations for Point Cloud Registration

Yue Wang, Justin M. Solomon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3523-3532

Point cloud registration is a key problem for computer vision applied to robotics, medical imaging, and other applications. This problem involves finding a rigid transformation from one point cloud into another so that they align. Iterative Closest Point (ICP) and its variants provide simple and easily-implemented iterative methods for this task, but these algorithms can converge to spurious local optima. To address local optima and other difficulties in the ICP pipeline, we propose a learning-based method, titled Deep Closest Point (DCP), inspired by recent techniques in computer vision and natural language processing. Our model consists of three parts: a point cloud embedding network, an attention-based module combined with a pointer generation layer to approximate combinatorial matching, and a differentiable singular value decomposition (SVD) layer to extract the final rigid transformation. We train our model end-to-end on the ModelNet40 dataset and show in several settings that it performs better than ICP, its variants (e.g., Go-ICP, FGR), and the recently-proposed learning-based method PointNetLK. Beyond providing a state-of-the-art registration technique, we evaluate the suitability of our learned features transferred to unseen objects. We also provide preliminary analysis of our learned model to help understand whether domain-specific and/or global features facilitate rigid registration.

****

Orientation-Aware Semantic Segmentation on Icosahedron Spheres

Chao Zhang, Stephan Liwicki, William Smith, Roberto Cipolla; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3533-3541

We address semantic segmentation on omnidirectional images, to leverage a holistic understanding of the surrounding scene for applications like autonomous driving systems. For the spherical domain, several methods recently adopt an icosahedron mesh, but systems are typically rotation invariant or require significant memory and parameters, thus enabling execution only at very low resolutions. In our work, we propose an orientation-aware CNN framework for the icosahedron mesh. Our representation allows for fast network operations, as our design simplifies to standard network operations of classical CNNs, but under consideration of north-aligned kernel convolutions for features on the sphere. We implement our representation and demonstrate its memory efficiency up-to a level-8 resolution mesh (equivalent to 640 x 1024 equirectangular images). Finally, since our kernels operate on the tangent of the sphere, standard feature weights, pretrained on perspective data, can be directly transferred with only small need for weight refinement. In our evaluation our orientation-aware CNN becomes a new state of the art for the recent 2D3DS dataset, and our Omni-SYNTHIA version of SYNTHIA. Rotation invariant classification and segmentation tasks are additionally presented for comparison to prior art.

****

Differentiable Learning-to-Group Channels via Groupable Convolutional Neural Networks

Zhaoyang Zhang, Jingyu Li, Wenqi Shao, Zhanglin Peng, Ruimao Zhang, Xiaogang Wang, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3542-3551

Group convolution, which divides the channels of ConvNets into groups, has achieved impressive improvement over the regular convolution operation. However, existing models, e.g. ResNext, still suffers from the sub-optimal performance due to manually defining the number of groups as a constant over all of the layers. To

ward addressing this issue, we present Groupable ConvNet (GroupNet) built by using a novel dynamic grouping convolution (DGConv) operation, which is able to learn the number of groups in an end-to-end manner. The proposed approach has several appealing benefits. (1) DGConv provides a unified convolution representation and covers many existing convolution operations such as regular dense convolution, group convolution, and depthwise convolution. (2) DGConv is a differentiable and flexible operation which learns to perform various convolutions from training data. (3) GroupNet trained with DGConv learns different number of groups for different convolution layers. Extensive experiments demonstrate that GroupNet outperforms its counterparts such as ResNet and ResNeXt in terms of accuracy and computational complexity. We also present introspection and reproducibility study, for the first time, showing the learning dynamics of training group numbers.
****

HarDNet: A Low Memory Traffic Network

Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, Youn-Long Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3552-3561

State-of-the-art neural network architectures such as ResNet, MobileNet, and DenseNet have achieved outstanding accuracy over low MACs and small model size counterparts. However, these metrics might not be accurate for predicting the inference time. We suggest that memory traffic for accessing intermediate feature maps can be a factor dominating the inference latency, especially in such tasks as real-time object detection and semantic segmentation of high-resolution video. We propose a Harmonic Densely Connected Network to achieve high efficiency in terms of both low MACs and memory traffic. The new network achieves 35%, 36%, 30%, 32%, and 45% inference time reduction compared with FC-DenseNet-103, DenseNet-264, ResNet-50, ResNet-152, and SSD-VGG, respectively. We use tools including Nvidia profiler and ARM Scale-Sim to measure the memory traffic and verify that the inference latency is indeed proportional to the memory traffic consumption and the proposed network consumes low memory traffic. We conclude that one should take memory traffic into consideration when designing neural network architectures for high-resolution applications at the edge.
****

Dynamic Multi-Scale Filters for Semantic Segmentation

Junjun He, Zhongying Deng, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3562-3572

Multi-scale representation provides an effective way to address scale variation of objects and stuff in semantic segmentation. Previous works construct multi-scale representation by utilizing different filter sizes, expanding filter sizes with dilated filters or pooling grids, and the parameters of these filters are fixed after training. These methods often suffer from heavy computational cost or have more parameters, and are not adaptive to the input image during inference. To address these problems, this paper proposes a Dynamic Multi-scale Network (DMNet) to adaptively capture multi-scale contents for predicting pixel-level semantic labels. DMNet is composed of multiple Dynamic Convolutional Modules (DCMs) arranged in parallel, each of which exploits context-aware filters to estimate semantic representation for a specific scale. The outputs of multiple DCMs are further integrated for final segmentation. We conduct extensive experiments to evaluate our DMNet on three challenging semantic segmentation and scene parsing datasets, PASCAL VOC 2012, Pascal-Context, and ADE20K. DMNet achieves a new record 84.4% mIoU on PASCAL VOC 2012 test set without MS COCO pre-trained and post-processing, and also obtains state-of-the-art performance on Pascal-Context and ADE20K.
****

Online Model Distillation for Efficient Video Inference

Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, Kayvon Fatahalian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3573-3582

High-quality computer vision models typically address the problem of understanding the general distribution of real-world images. However, most cameras observe only a very small fraction of this distribution. This offers the possibility of achieving more efficient inference by specializing compact, low-cost models to the specific distribution of frames observed by a single camera. In this paper, we employ the technique of model distillation (supervising a low-cost student model using the output of a high-cost teacher) to specialize accurate, low-cost semantic segmentation models to a target video stream. Rather than learn a specialized student model on offline data from the video stream, we train the student in an online fashion on the live video, intermittently running the teacher to provide a target for learning. Online model distillation yields semantic segmentation models that closely approximate their Mask R-CNN teacher with 7 to 17xlower inference runtime cost (11 to 26xin FLOPs), even when the target video's distribution is non-stationary. Our method requires no offline pretraining on the target video stream, achieves higher accuracy and lower cost than solutions based on flow or video object segmentation, and can exhibit better temporal stability than the original teacher. We also provide a new video dataset for evaluating the efficiency of inference over long running video streams.

****

Rethinking Zero-Shot Learning: A Conditional Visual Classification Perspective
Kai Li, Martin Renqiang Min, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3583-3592

Zero-shot learning (ZSL) aims to recognize instances of unseen classes solely based on the semantic descriptions of the classes. Existing algorithms usually formulate it as a semantic-visual correspondence problem, by learning mappings from one feature space to the other. Despite being reasonable, previous approaches essentially discard the highly precious discriminative power of visual features in an implicit way, and thus produce undesirable results. We instead reformulate ZSL as a conditioned visual classification problem, i.e., classifying visual features based on the classifiers learned from the semantic descriptions. With this reformulation, we develop algorithms targeting various ZSL settings: For the conventional setting, we propose to train a deep neural network that directly generates visual feature classifiers from the semantic attributes with an episode-based training scheme; For the generalized setting, we concatenate the learned highly discriminative classifiers for seen classes and the generated classifiers for unseen classes to classify visual features of all classes; For the transductive setting, we exploit unlabeled data to effectively calibrate the classifier generator using a novel learning-without-forgetting self-training mechanism and guide the process by a robust generalized cross-entropy loss. Extensive experiments show that our proposed algorithms significantly outperform state-of-the-art methods by large margins on most benchmark datasets in all the ZSL settings.

****

Task-Driven Modular Networks for Zero-Shot Compositional Learning
Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, Marc'Aurelio Ranzato; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3593-3602

One of the hallmarks of human intelligence is the ability to compose learned knowledge into novel concepts which can be recognized without a single training example. In contrast, current state-of-the-art methods require hundreds of training examples for each possible category to build reliable and accurate classifiers. To alleviate this striking difference in efficiency, we propose a task-driven modular architecture for compositional reasoning and sample efficient learning. Our architecture consists of a set of neural network modules, which are small fully connected layers operating in semantic concept space. These modules are configured through a gating function conditioned on the task to produce features representing the compatibility between the input image and the concept under consideration. This enables us to express tasks as a combination of sub-tasks and to generalize to unseen categories by reweighting a set of small modules. Furthermore, the network can be trained efficiently as it is fully differentiable and its modules operate on small sub-spaces. We focus our study on the problem of compositional zero-shot classification of object-attribute categories. We show in our experiments that current evaluation metrics are flawed as they only consider unse en object-attribute pairs. When extending the evaluation to the generalized setting which accounts also for pairs seen during training, we discover that naive baseline methods perform similarly or better than current approaches. However, our modular network is able to outperform all existing approaches on two widely-used benchmark datasets.

****

Transductive Episodic-Wise Adaptive Metric for Few-Shot Learning

Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3603-3612

Few-shot learning, which aims at extracting new concepts rapidly from extremely few examples of novel classes, has been featured into the meta-learning paradigm recently. Yet, the key challenge of how to learn a generalizable classifier with the capability of adapting to specific tasks with severely limited data still remains in this domain. To this end, we propose a Transductive Episodic-wise Adaptive Metric (TEAM) framework for few-shot learning, by integrating the meta-learning paradigm with both deep metric learning and transductive inference. With exploring the pairwise constraints and regularization prior within each task, we explicitly formulate the adaptation procedure into a standard semi-definite programming problem. By solving the problem with its closed-form solution on the fly with the setup of transduction, our approach efficiently tailors an episodic-wise metric for each task to adapt all features from a shared task-agnostic embedding space into a more discriminative task-specific metric space. Moreover, we further leverage an attention-based bi-directional similarity strategy for extracting the more robust relationship between queries and prototypes. Extensive experiments on three benchmark datasets show that our framework is superior to other existing approaches and achieves the state-of-the-art performance in the few-shot literature.

****

Deep Multiple-Attribute-Perceived Network for Real-World Texture Recognition

Wei Zhai, Yang Cao, Jing Zhang, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3613-3622

Texture recognition is a challenging visual task as multiple perceptual attributes may be perceived from the same texture image when combined with different spatial context. Some recent works building upon Convolutional Neural Network (CNN) incorporate feature encoding with orderless aggregating to provide invariance to spatial layouts. However, these existing methods ignore visual texture attributes, which are important cues for describing the real-world texture images, resulting in incomplete description and inaccurate recognition. To address this problem, we propose a novel deep Multiple-Attribute-Perceived Network (MAP-Net) by progressively learning visual texture attributes in a mutually reinforced manner. Specifically, a multi-branch network architecture is devised, in which cascaded global contexts are learned by introducing similarity constraint at each branch, and leveraged as guidance of spatial feature encoding at next branch through an attribute transfer scheme. To enhance the modeling capability of spatial transformation, a deformable pooling strategy is introduced to augment the spatial sampling with adaptive offsets to the global context, leading to perceive new visual attributes. An attribute fusion module is then introduced to jointly utilize the perceived visual attributes and the abstracted semantic concepts at each branch. Experimental results on the five most challenging texture recognition datasets have demonstrated the superiority of the proposed model against the state-of-the-arts.

****

RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment

Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, Zengguang Hou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3623-3632

RGB-Infrared (IR) person re-identification is an important and challenging task due to large cross-modality variations between RGB and IR images. Most conventio nal approaches aim to bridge the cross-modality gap with feature alignment by feature representation learning. Different from existing methods, in this paper, we propose a novel and end-to-end Alignment Generative Adversarial Network (AlignGAN) for the RGB-IR RE-ID task. The proposed model enjoys several merits. First, it can exploit pixel alignment and feature alignment jointly. To the best of our knowledge, this is the first work to model the two alignment strategies jointly for the RGB-IR RE-ID problem. Second, the proposed model consists of a pixel generator, a feature generator and a joint discriminator. By playing a min-max game among the three components, our model is able to not only alleviate the cross-modality and intra-modality variations, but also learn identity-consistent features. Extensive experimental results on two standard benchmarks demonstrate that the proposed model performs favourably against state-of-the-art methods. Especially, on SYSU-MM01 dataset, our model can achieve an absolute gain of 15.4% and 12.9% in terms of Rank-1 and mAP.

****

EvalNorm: Estimating Batch Normalization Statistics for Evaluation

Saurabh Singh, Abhinav Shrivastava; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3633-3641

Batch normalization (BN) has been very effective for deep learning and is widely used. However, when training with small minibatches, models using BN exhibit a significant degradation in performance. In this paper we study this peculiar behavior of BN to gain a better understanding of the problem, and identify a cause. We propose `EvalNorm' to address the issue by estimating corrected normalization statistics to use for BN during evaluation. EvalNorm supports online estimation of the corrected statistics while the model is being trained, and does not affect the training scheme of the model. As a result, EvalNorm can also be used with existing pre-trained models allowing them to benefit from our method. EvalNorm yields large gains for models trained with smaller batches. Our experiments show that EvalNorm performs 6.18% (absolute) better than vanilla BN for a batchsize of 2 on ImageNet validation set and from 1.5 to 7.0 points (absolute) gain on the COCO object detection benchmark across a variety of setups.

****

Beyond Human Parts: Dual Part-Aligned Representations for Person Re-Identification

Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, Kai Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3642-3651

Person re-identification is a challenging task due to various complex factors. Recent studies have attempted to integrate human parsing results or externally defined attributes to help capture human parts or important object regions. On the other hand, there still exist many useful contextual cues that do not fall into the scope of predefined human parts or attributes. In this paper, we address the missed contextual cues by exploiting both the accurate human parts and the coarse non-human parts. In our implementation, we apply a human parsing model to extract the binary human part masks and a self-attention mechanism to capture the soft latent (non-human) part masks. We verify the effectiveness of our approach with new state-of-the-art performance on three challenging benchmarks: Market-1501, DukeMTMC-reID and CUHK03. Our implementation is available at https://github.com/ggjy/P2Net.pytorch.

****

Person Search by Text Attribute Query As Zero-Shot Learning

Qi Dong, Shaogang Gong, Xiatian Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3652-3661

Existing person search methods predominantly assume the availability of at least one-shot imagery sample of the queried person. This assumption is limited in circumstances where only a brief textual (or verbal) description of the target person is available. In this work, we present a deep learning method for attribute text description based person search without any query imagery. Whilst conventional cross-modality matching methods, such as global visual-textual embedding based zero-shot learning and local individual attribute recognition, are functional ly applicable, they are limited by several assumptions invalid to person search in deployment scale, data quality, and/or category name semantics. We overcome these issues by formulating an Attribute-Image Hierarchical Matching (AIHM) model. It is able to more reliably match text attribute descriptions with noisy surveillance person images by jointly learning global category-level and local attribute-level textual-visual embedding as well as matching. Extensive evaluations demonstrate the superiority of our AIHM model over a wide variety of state-of-the-art methods on three publicly available attribute labelled surveillance person search benchmarks: Market-1501, DukeMTMC, and PA100K.

****

Semantic-Aware Knowledge Preservation for Zero-Shot Sketch-Based Image Retrieval

Qing Liu, Lingxi Xie, Huiyu Wang, Alan L. Yuille; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3662-3671

Sketch-based image retrieval (SBIR) is widely recognized as an important vision problem which implies a wide range of real-world applications. Recently, research interests arise in solving this problem under the more realistic and challenging setting of zero-shot learning. In this paper, we investigate this problem from the viewpoint of domain adaptation which we show is critical in improving feature embedding in the zero-shot scenario. Based on a framework which starts with a pre-trained model on ImageNet and fine-tunes it on the training set of SBIR benchmark, we advocate the importance of preserving previously acquired knowledge, e.g., the rich discriminative features learned from ImageNet, to improve the model's transfer ability. For this purpose, we design an approach named Semantic-Aware Knowledge prEservation (SAKE), which fine-tunes the pre-trained model in an economical way and leverages semantic information, e.g., inter-class relationship, to achieve the goal of knowledge preservation. Zero-shot experiments on two extended SBIR datasets, TU-Berlin and Sketchy, verify the superior performance of our approach. Extensive diagnostic experiments validate that knowledge preserved benefits SBIR in zero-shot settings, as a large fraction of the performance gain is from the more properly structured feature embedding for photo images.

****

Active Learning for Deep Detection Neural Networks

Hamed H. Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, Antonio M. Lopez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3672-3680

The cost of drawing object bounding boxes (i.e. labeling) for millions of images is prohibitively high. For instance, labeling pedestrians in a regular urban image could take 35 seconds on average. Active learning aims to reduce the cost of labeling by selecting only those images that are informative to improve the detection network accuracy. In this paper, we propose a method to perform active learning of object detectors based on convolutional neural networks. We propose a new image-level scoring process to rank unlabeled images for their automatic selection, which clearly outperforms classical scores. The proposed method can be applied to videos and sets of still images. In the former case, temporal selection rules can complement our scoring process. As a relevant use case, we extensively study the performance of our method on the task of pedestrian detection. Overall, the experiments show that the proposed method performs better than random selection.

****

One-Shot Neural Architecture Search via Self-Evaluated Template Network

Xuanyi Dong, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3681-3690

Neural architecture search (NAS) aims to automate the search procedure of architecture instead of manual design. Even if recent NAS approaches finish the search within days, lengthy training is still required for a specific architecture candidate to get the parameters for its accurate evaluation. Recently one-shot NAS methods are proposed to largely squeeze the tedious training process by sharing parameters across candidates. In this way, the parameters for each candidate can be directly extracted from the shared parameters instead of training them from scratch. However, they have no sense of which candidate will perform better unti

l evaluation so that the candidates to evaluate are randomly sampled and the top-1 candidate is considered the best. In this paper, we propose a Self-Evaluated Template Network (SETN) to improve the quality of the architecture candidates for evaluation so that it is more likely to cover competitive candidates. SETN consists of two components: (1) an evaluator, which learns to indicate the probability of each individual architecture being likely to have a lower validation loss. The candidates for evaluation can thus be selectively sampled according to this evaluator. (2) a template network, which shares parameters among all candidates to amortize the training cost of generated candidates. In experiments, the architecture found by SETN achieves the state-of-the-art performance on CIFAR and ImageNet benchmarks within comparable computation costs.
****

Batch DropBlock Network for Person Re-Identification and Beyond
Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, Ping Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3691-3701
Since the person re-identification task often suffers from the problem of pose changes and occlusions, some attentive local features are often suppressed when training CNNs. In this paper, we propose the Batch DropBlock (BDB) Network which is a two branch network composed of a conventional ResNet-50 as the global branch and a feature dropping branch.The global branch encodes the global salient representations.Meanwhile, the feature dropping branch consists of an attentive feature learning module called Batch DropBlock, which randomly drops the same region of all input feature maps in a batch to reinforce the attentive feature learning of local regions.The network then concatenates features from both branches and provides a more comprehensive and spatially distributed feature representation. Albeit simple, our method achieves state-of-the-art on person re-identification and it is also applicable to general metric learning tasks. For instance, we achieve 76.4% Rank-1 accuracy on the CUHK03-Detect dataset and 83.0% Recall-1 score on the Stanford Online Products dataset, outperforming the existed works by a large margin (more than 6%).
****

Omni-Scale Feature Learning for Person Re-Identification
Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3702-3712
As an instance-level recognition problem, person re-identification (ReID) relies on discriminative features, which not only capture different spatial scales but also encapsulate an arbitrary combination of multiple scales. We callse features of both homogeneous and heterogeneous scales omni-scale features. In this paper, a novel deep ReID CNN is designed, termed Omni-Scale Network (OSNet), for omni-scale feature learning. This is achieved by designing a residual block composed of multiple convolutional feature streams, each detecting features at a certain scale. Importantly, a novel unified aggregation gate is introduced to dynamically fuse multi-scale features with input-dependent channel-wise weights. To efficiently learn spatial-channel correlations and avoid overfitting, the building block uses both pointwise and depthwise convolutions. By stacking such blocks layer-by-layer, our OSNet is extremely lightweight and can be trained from scratch on existing ReID benchmarks. Despite its small model size, our OSNet achieves state-of-the-art performance on six person-ReID datasets. Code and models are available at: https://github.com/KaiyangZhou/deep-person-reid.
****

Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation
Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, Kaisheng Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3713-3722
Convolutional neural networks have been widely deployed in various application scenarios. In order to extend the applications' boundaries to some accuracy-crucial domains, researchers have been investigating approaches to boost accuracy through either deeper or wider network structures, which brings with them the expon

ential increment of the computational and storage cost, delaying the responding time. In this paper, we propose a general training framework named self distillation, which notably enhances the performance (accuracy) of convolutional neural networks through shrinking the size of the network rather than aggrandizing it. Different from traditional knowledge distillation - a knowledge transformation methodology among networks, which forces student neural networks to approximate the softmax layer outputs of pre-trained teacher neural networks, the proposed self distillation framework distills knowledge within network itself. The networks are firstly divided into several sections. Then the knowledge in the deeper portion of the networks is squeezed into the shallow ones. Experiments further prove the generalization of the proposed self distillation framework: enhancement of accuracy at average level is 2.65%, varying from 0.61% in ResNeXt as minimum to 4.07% in VGG19 as maximum. In addition, it can also provide flexibility of depth-wise scalable inference on resource-limited edge devices. Our codes have been released on github.

****

Diversity With Cooperation: Ensemble Methods for Few-Shot Classification

Nikita Dvornik, Cordelia Schmid, Julien Mairal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3723-3731

Few-shot classification consists of learning a predictive model that is able to effectively adapt to a new class, given only a few annotated samples. To solve this challenging problem, meta-learning has become a popular paradigm that advocates the ability to "learn to adapt". Recent works have shown, however, that simple learning strategies without meta-learning could be competitive. In this paper, we go a step further and show that by addressing the fundamental high-variance issue of few-shot learning classifiers, it is possible to significantly outperform current meta-learning techniques. Our approach consists of designing an ensemble of deep networks to leverage the variance of the classifiers, and introducing new strategies to encourage the networks to cooperate, while encouraging prediction diversity. Evaluation is conducted on the mini-ImageNet, tiered-ImageNet and CUB datasets, where we show that even a single network obtained by distillation yields state-of-the-art results.

****

Enhancing 2D Representation via Adjacent Views for 3D Shape Retrieval

Cheng Xu, Zhaoqun Li, Qiang Qiu, Biao Leng, Jingfei Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3732-3740

Multi-view shape descriptors obtained from various 2D images are commonly adopted in 3D shape retrieval. One major challenge is that significant shape information are discarded during 2D view rendering through projection. In this paper, we propose a convolutional neural network based method, CenterNet, to enhance each individual 2D view using its neighboring ones. By exploiting cross-view correlations, CenterNet learns how adjacent views can be maximally incorporated for an enhanced 2D representation to effectively describe shapes. We observe that a very small amount of, e.g., six, enhanced 2D views, are already sufficient for a panoramic shape description. Thus, by simply aggregating features from six enhanced 2D views, we arrive at a highly compact yet discriminative shape descriptor. The proposed shape descriptor significantly outperforms state-of-the-art 3D shape retrieval methods on the ModelNet and ShapeNetCore55 benchmarks, and also exhibits its robustness against object occlusion.

****

Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition

Kun Wei, Muli Yang, Hao Wang, Cheng Deng, Xianglong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3741-3749

Recognizing unseen attribute-object pairs never appearing in the training data is a challenging task, since an object often refers to a specific entity while an attribute is an abstract semantic description. Besides, attributes are highly correlated to objects, i.e., an attribute tends to describe different visual features of various objects. Existing methods mainly employ two classifiers to recog nize attribute and object separately, or simply simulate the composition of attribute and object, which ignore the inherent discrepancy and correlation between them. In this paper, we propose a novel adversarial fine-grained composition learning model for unseen attribute-object pair recognition. Considering their inherent discrepancy, we leverage multi-scale feature integration to capture discriminative fine-grained features from a given image. Besides, we devise a quintuplet loss to depict more accurate correlations between attributes and objects. Adversarial learning is employed to model the discrepancy and correlations among attributes and objects. Extensive experiments on two challenging benchmarks indicate that our method consistently outperforms state-of-the-art competitors by a large margin.

****

Auto-ReID: Searching for a Part-Aware ConvNet for Person Re-Identification

Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3750-3759

Prevailing deep convolutional neural networks (CNNs) for person re-IDentification (reID) are usually built upon ResNet or VGG backbones, which were originally designed for classification. Because reID is different from classification, the architecture should be modified accordingly. We propose to automatically search for a CNN architecture that is specifically suitable for the reID task. There are three aspects to be tackled. First, body structural information plays an important role in reID but it is not encoded in backbones. Second, Neural Architecture Search (NAS) automates the process of architecture design without human effort, but no existing NAS methods incorporate the structure information of input images. Third, reID is essentially a retrieval task but current NAS algorithms are merely designed for classification. To solve these problems, we propose a retrieval-based search algorithm over a specifically designed reID search space, named Auto-ReID. Our Auto-ReID enables the automated approach to find an efficient and effective CNN architecture for reID. Extensive experiments demonstrate that the searched architecture achieves state-of-the-art performance while reducing 50% parameters and 53% FLOPs compared to others.

****

Second-Order Non-Local Attention Networks for Person Re-Identification

Bryan (Ning) Xia, Yuan Gong, Yizhe Zhang, Christian Poellabauer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3760-3769

Recent efforts have shown promising results for person re-identification by designing part-based architectures to allow a neural network to learn discriminative representations from semantically coherent parts. Some efforts use soft attention to reallocate distant outliers to their most similar parts, while others adjust part granularity to incorporate more distant positions for learning the relationships. Others seek to generalize part-based methods by introducing a dropout mechanism on consecutive regions of the feature map to enhance distant region relationships. However, only few prior efforts model the distant or non-local positions of the feature map directly for the person re-ID task. In this paper, we propose a novel attention mechanism to directly model long-range relationships via a second-order feature statistics. When combined with a generalized DropBlock module, our method performs equally to or better than state-of-the-art results for mainstream person re-identification datasets, including Market1501, CUHK03, and DukeMTMC-reID.

****

Fast Computation of Content-Sensitive Superpixels and Supervoxels Using Q-Distances

Zipeng Ye, Ran Yi, Minjing Yu, Yong-Jin Liu, Ying He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3770-3779

State-of-the-art researches model the data of images and videos as low-dimensional manifolds and generate superpixels/supervoxels in a content-sensitive way, which is achieved by computing geodesic centroidal Voronoi tessellation (GCVT) on manifolds. However, computing exact GCVTs is slow due to computationally expensive geodesic distances. In this paper, we propose a much faster queue-based graph distance (called q-distance). Our key idea is that for manifold regions in which q-distances are different from geodesic distances, GCVT is prone to placing more generators in them, and therefore after few iterations, the q-distance-induced tessellation is an exact GCVT. This idea works well in practice and we also prove it theoretically under moderate assumption. Our method is simple and easy to implement. It runs 6-8 times faster than state-of-the-art GCVT computation, and has an optimal approximation ratio O(1) and a linear time complexity O(N) for N-pixel images or N-voxel videos. A thorough evaluation of 31 superpixel methods on five image datasets and 8 supervoxel methods on four video datasets shows that our method consistently achieves the best over-segmentation accuracy. We also demonstrate the advantage of our method on one image and two video applications.

****

Progressive-X: Efficient, Anytime, Multi-Model Fitting Algorithm

Daniel Barath, Jiri Matas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3780-3788

The Progressive-X algorithm, Prog-X in short, is proposed for geometric multi-model fitting. The method interleaves sampling and consolidation of the current data interpretation via repetitive hypothesis proposal, fast rejection, and integration of the new hypothesis into the kept instance set by labeling energy minimization. Due to exploring the data progressively, the method has several beneficial properties compared with the state-of-the-art. First, a clear criterion, adopted from RANSAC, controls the termination and stops the algorithm when the probability of finding a new model with a reasonable number of inliers falls below a threshold. Second, Prog-X is an any-time algorithm. Thus, whenever is interrupted, e.g. due to a time limit, the returned instances cover real and, likely, the most dominant ones. The method is superior to the state-of-the-art in terms of accuracy in both synthetic experiments and on publicly available real-world datasets for homography, two-view motion, and motion segmentation.

****

Structured Modeling of Joint Deep Feature and Prediction Refinement for Salient Object Detection

Yingyue Xu, Dan Xu, Xiaopeng Hong, Wanli Ouyang, Rongrong Ji, Min Xu, Guoying Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3789-3798

Recent saliency models extensively explore to incorporate multi-scale contextual information from Convolutional Neural Networks (CNNs). Besides direct fusion strategies, many approaches introduce message-passing to enhance CNN features or predictions. However, the messages are mainly transmitted in two ways, by feature-to-feature passing, and by prediction-to-prediction passing. In this paper, we add message-passing between features and predictions and propose a deep unified CRF saliency model . We design a novel cascade CRFs architecture with CNN to jointly refine deep features and predictions at each scale and progressively compute a final refined saliency map. We formulate the CRF graphical model that involves message-passing of feature-feature, feature-prediction, and prediction-prediction, from the coarse scale to the finer scale, to update the features and the corresponding predictions. Also, we formulate the mean-field updates for joint end-to-end model training with CNN through back propagation. The proposed deep unified CRF saliency model is evaluated over six datasets and shows highly competitive performance among the state of the arts.

****

Selectivity or Invariance: Boundary-Aware Salient Object Detection

Jinming Su, Jia Li, Yu Zhang, Changqun Xia, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3799-3808

Typically, a salient object detection (SOD) model faces opposite requirements in processing object interiors and boundaries. The features of interiors should be invariant to strong appearance change so as to pop-out the salient object as a whole, while the features of boundaries should be selective to slight appearance change to distinguish salient objects and background. To address this selectivity-invariance dilemma, we propose a novel boundary-aware network with successive dilation for image-based SOD. In this network, the feature selectivity at boundaries is enhanced by incorporating a boundary localization stream, while the feature invariance at interiors is guaranteed with a complex interior perception stream. Moreover, a transition compensation stream is adopted to amend the probable failures in transitional regions between interiors and boundaries. In particular, an integrated successive dilation module is proposed to enhance the feature invariance at interiors and transitional regions. Extensive experiments on six datasets show that the proposed approach outperforms 16 state-of-the-art methods.

****

Online Unsupervised Learning of the 3D Kinematic Structure of Arbitrary Rigid Bodies

Urbano Miguel Nunes, Yiannis Demiris; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3809-3817

This work addresses the problem of 3D kinematic structure learning of arbitrary articulated rigid bodies from RGB-D data sequences. Typically, this problem is addressed by offline methods that process a batch of frames, assuming that complete point trajectories are available. However, this approach is not feasible when considering scenarios that require continuity and fluidity, for instance, human-robot interaction. In contrast, we propose to tackle this problem in an online unsupervised fashion, by recursively maintaining the metric distance of the scene's 3D structure, while achieving real-time performance. The influence of noise is mitigated by building a similarity measure based on a linear embedding representation and incorporating this representation into the original metric distance. The kinematic structure is then estimated based on a combination of implicit motion and spatial properties. The proposed approach achieves competitive performance both quantitatively and qualitatively in terms of estimation accuracy, even compared to offline methods.

****

Few-Shot Generalization for Single-Image 3D Reconstruction via Priors

Bram Wallace, Bharath Hariharan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3818-3827

Recent work on single-view 3D reconstruction shows impressive results, but has been restricted to a few fixed categories where extensive training data is available. The problem of generalizing these models to new classes with limited training data is largely open. To address this problem, we present a new model architecture that reframes single-view 3D reconstruction as learnt, category agnostic refinement of a provided, category-specific prior. The provided prior shape for a novel class can be obtained from as few as one 3D shape from this class. Our model can start reconstructing objects from the novel class using this prior without seeing any training image for this class and without any retraining. Our model outperforms category-agnostic baselines and remains competitive with more sophisticated baselines that finetune on the novel categories. Additionally, our network is capable of improving the reconstruction given multiple views despite not being trained on task of multi-view reconstruction.

****

Digging Into Self-Supervised Monocular Depth Estimation

Clement Godard, Oisin Mac Aodha, Michael Firman, Gabriel J. Brostow; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3828-3838

Per-pixel ground-truth depth data is challenging to acquire at scale. To overcome this limitation, self-supervised learning has emerged as a promising alternative for training models to perform monocular depth estimation. In this paper, we propose a set of improvements, which together result in both quantitatively and qualitatively improved depth maps compared to competing self-supervised methods. Research on self-supervised monocular training usually explores increasingly complex architectures, loss functions, and image formation models, all of which have recently helped to close the gap with fully-supervised methods. We show that a surprisingly simple model, and associated design choices, lead to superior predictions. In particular, we propose (i) a minimum reprojection loss, designed to robustly handle occlusions, (ii) a full-resolution multi-scale sampling method that reduces visual artifacts, and (iii) an auto-masking loss to ignore training pixels that violate camera motion assumptions. We demonstrate the effectiveness of each component in isolation, and show high quality, state-of-the-art results on the KITTI benchmark.

****

Learning Object-Specific Distance From a Monocular Image

Jing Zhu, Yi Fang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3839-3848

Environment perception, including object detection and distance estimation, is one of the most crucial tasks for autonomous driving. Many attentions have been paid on the object detection task, but distance estimation only arouse few interests in the computer vision community. Observing that the traditional inverse perspective mapping algorithm performs poorly for objects far away from the camera or on the curved road, in this paper, we address the challenging distance estimation problem by developing the first end-to-end learning-based model to directly predict distances for given objects in the images. Besides the introduction of a learning-based base model, we further design an enhanced model with a keypoint regressor, where a projection loss is defined to enforce a better distance estimation, especially for objects close to the camera. To facilitate the research on this task, we construct the extented KITTI and nuScenes (mini) object detection datasets with a distance for each object. Our experiments demonstrate that our proposed methods outperform alternative approaches (e.g., the traditional IPM, SVR) on object-specific distance estimation, particularly for the challenging cases that objects are on a curved road. Moreover, the performance margin implies the effectiveness of our enhanced method.

****

Unsupervised 3D Reconstruction Networks

Geonho Cha, Minsik Lee, Songhwai Oh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3849-3858

In this paper, we propose 3D unsupervised reconstruction networks (3D-URN), which reconstruct the 3D structures of instances in a given object category from their 2D feature points under an orthographic camera model. 3D-URN consists of a 3D shape reconstructor and a rotation estimator, which are trained in a fully-unsupervised manner incorporating the proposed unsupervised loss functions. The role of the 3D shape reconstructor is to reconstruct the 3D shape of an instance from its 2D feature points, and the rotation estimator infers the camera pose. After training, 3D-URN can infer the 3D structure of an unseen instance in the same category, which is not possible in the conventional schemes of non-rigid structure from motion and structure from category. The experimental result shows the state-of-the-art performance, which demonstrates the effectiveness of the proposed method.

****

3D Point Cloud Generative Adversarial Network Based on Tree Structured Graph Convolutions

Dong Wook Shu, Sung Woo Park, Junseok Kwon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3859-3868

In this paper, we propose a novel generative adversarial network (GAN) for 3D point clouds generation, which is called tree-GAN. To achieve state-of-the-art performance for multi-class 3D point cloud generation, a tree-structured graph convolution network (TreeGCN) is introduced as a generator for tree-GAN. Because TreeGCN performs graph convolutions within a tree, it can use ancestor information to boost the representation power for features. To evaluate GANs for 3D point clouds accurately, we develop a novel evaluation metric called Frechet point cloud distance (FPD). Experimental results demonstrate that the proposed tree-GAN outperforms state-of-the-art GANs in terms of both conventional metrics and FPD, and can generate point clouds for different semantic parts without prior knowledge.

****

Visualization of Convolutional Neural Networks for Monocular Depth Estimation

Junjie Hu, Yan Zhang, Takayuki Okatani; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2019, pp. 3869-3878

Recently, convolutional neural networks (CNNs) have shown great success on the task of monocular depth estimation. A fundamental yet unanswered question is: how CNNs can infer depth from a single image. Toward answering this question, we consider visualization of inference of a CNN by identifying relevant pixels of an input image to depth estimation. We formulate it as an optimization problem of identifying the smallest number of image pixels from which the CNN can estimate a depth map with the minimum difference from the estimate from the entire image. To cope with a difficulty with optimization through a deep CNN, we propose to use another network to predict those relevant image pixels in a forward computation. In our experiments, we first show the effectiveness of this approach, and then apply it to different depth estimation networks on indoor and outdoor scene datasets. The results provide several findings that help exploration of the above question.

****

Co-Separating Sounds of Visual Objects

Ruohan Gao, Kristen Grauman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3879-3888

Learning how objects sound from video is challenging, since they often heavily overlap in a single audio channel. Current methods for visually-guided audio source separation sidestep the issue by training with artificially mixed video clips, but this puts unwieldy restrictions on training data collection and may even prevent learning the properties of "true" mixed sounds. We introduce a co-separation training paradigm that permits learning object-level sounds from unlabeled multi-source videos. Our novel training objective requires that the deep neural network's separated audio for similar-looking objects be consistently identifiable, while simultaneously reproducing accurate video-level audio tracks for each source training pair. Our approach disentangles sounds in realistic test videos, even in cases where an object was not observed individually during training. We obtain state-of-the-art results on visually-guided audio source separation and audio denoising for the MUSIC, AudioSet, and AV-Bench datasets.

****

BMN: Boundary-Matching Network for Temporal Action Proposal Generation

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, Shilei Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3889-3898

Temporal action proposal generation is an challenging and promising task which aims to locate temporal regions in real-world videos where action or event may occur. Current bottom-up proposal generation methods can generate proposals with precise boundary, but cannot efficiently generate adequately reliable confidence scores for retrieving proposals. To address these difficulties, we introduce the Boundary-Matching (BM) mechanism to evaluate confidence scores of densely distributed proposals, which denote a proposal as a matching pair of starting and ending boundaries and combine all densely distributed BM pairs into the BM confidence map. Based on BM mechanism, we propose an effective, efficient and end-to-end proposal generation method, named Boundary-Matching Network (BMN), which generates proposals with precise temporal boundaries as well as reliable confidence scores simultaneously. The two-branches of BMN are jointly trained in an unified framework. We conduct experiments on two challenging datasets: THUMOS-14 and ActivityNet-1.3, where BMN shows significant performance improvement with remarkable efficiency and generalizability. Further, combining with existing action classifier, BMN can achieve state-of-the-art temporal action detection performance.

****

Weakly Supervised Temporal Action Localization Through Contrast Based Evaluation Networks

Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, Gang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3899-3908

Weakly-supervised temporal action localization (WS-TAL) is a promising but challenging task with only video-level action categorical labels available during training. Without requiring temporal action boundary annotations in training data,

WS-TAL could possibly exploit automatically retrieved video tags as video-level labels. However, such coarse video-level supervision inevitably incurs confusions, especially in untrimmed videos containing multiple action instances. To address this challenge, we propose the Contrast-based Localization EvaluAtioN Network (CleanNet) with our new action proposal evaluator, which provides pseudo-supervision by leveraging the temporal contrast in snippet-level action classification predictions. Essentially, the new action proposal evaluator enforces an additional temporal contrast constraint so that high-evaluation-score action proposals are more likely to coincide with true action instances. Moreover, the new action localization module is an integral part of CleanNet which enables end-to-end training. This is in contrast to many existing WS-TAL methods where action localization is merely a post-processing step. Experiments on THUMOS14 and ActivityNet datasets validate the efficacy of CleanNet against existing state-ofthe- art WS-TAL algorithms.
****

Progressive Sparse Local Attention for Video Object Detection
Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, Chunhong Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3909-3918

Transferring image-based object detectors to the domain of videos remains a challenging problem. Previous efforts mostly exploit optical flow to propagate features across frames, aiming to achieve a good trade-off between accuracy and efficiency. However, introducing an extra model to estimate optical flow can significantly increase the overall model size. The gap between optical flow and high-level features can also hinder it from establishing spatial correspondence accurately. Instead of relying on optical flow, this paper proposes a novel module called Progressive Sparse Local Attention (PSLA), which establishes the spatial correspondence between features across frames in a local region with progressively sparser stride and uses the correspondence to propagate features. Based on PSLA, Recursive Feature Updating (RFU) and Dense Feature Transforming (DenseFT) are proposed to model temporal appearance and enrich feature representation respectively in a novel video object detection framework. Experiments on ImageNet VID show that our method achieves the best accuracy compared to existing methods with smaller model size and acceptable runtime speed.
****

Reasoning About Human-Object Interactions Through Dual Attention Networks
Tete Xiao, Quanfu Fan, Dan Gutfreund, Mathew Monfort, Aude Oliva, Bolei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3919-3928

Objects are entities we act upon, where the functionality of an object is determined by how we interact with it. In this work we propose a Dual Attention Network model which reasons about human-object interactions. The dual-attentional framework weights the important features for objects and actions respectively. As a result, the recognition of objects and actions mutually benefit each other. The proposed model shows competitive classification performance on the human-object interaction dataset Something-Something. Besides, it can perform weak spatiotemporal localization and affordance segmentation, despite being trained only with video-level labels. The model not only finds when an action is happening and which object is being manipulated, but also identifies which part of the object is being interacted with.
****

DMM-Net: Differentiable Mask-Matching Network for Video Object Segmentation
Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3929-3938

In this paper, we propose the differentiable mask-matching network (DMM-Net) for solving the video object segmentation problem where the initial object masks are provided. Relying on the Mask R-CNN backbone, we extract mask proposals per frame and formulate the matching between object templates and proposals as a linear assignment problem where thA heading inside a blocke cost matrix is predicted by a deep convolutional neural network. We propose a differentiable matching layer which unrolls a projected gradient descent algorithm in which the projection step exploits the Dykstra's algorithm. We prove that under mild conditions, the matching is guaranteed to converge to the optimal one. In practice, it achieves similar performance compared to the Hungarian algorithm during inference. Meanwhile, we can back-propagate through it to learn the cost matrix. After matching, a U-Net style architecture is exploited to refine the matched mask per time step. On DAVIS 2017 dataset, DMM-Net achieves the best performance without online learning on the first frames and the 2nd best with it. Without any fine-tuning, DMM-Net performs comparably to state-of-the-art methods on SegTrack v2 dataset. At last, our differentiable matching layer is very simple to implement; we attach the PyTorch code in the supplementary material which is less than 50 lines long.

****

Asymmetric Cross-Guided Attention Network for Actor and Action Video Segmentation From Natural Language Query

Hao Wang, Cheng Deng, Junchi Yan, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3939-3948

Actor and action video segmentation from natural language query aims to selectively segment the actor and its action in a video based on an input textual description. Previous works mostly focus on learning simple correlation between two heterogeneous features of vision and language via dynamic convolution or fully convolutional classification. However, they ignore the linguistic variation of natural language query and have difficulty in modeling global visual context, which leads to unsatisfactory segmentation performance. To address these issues, we propose an asymmetric cross-guided attention network for actor and action video segmentation from natural language query. Specifically, we frame an asymmetric cross-guided attention network, which consists of vision guided language attention to reduce the linguistic variation of input query and language guided vision attention to incorporate query-focused global visual context simultaneously. Moreover, we adopt multi-resolution fusion scheme and weighted loss for foreground and background pixels to obtain further performance improvement. Extensive experiments on Actor-Action Dataset Sentences and J-HMDB Sentences show that our proposed approach notably outperforms state-of-the-art methods.

****

AGSS-VOS: Attention Guided Single-Shot Video Object Segmentation

Huaijia Lin, Xiaojuan Qi, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3949-3957

Most video object segmentation approaches process objects separately. This incurs high computational cost when multiple objects exist. In this paper, we propose AGSS-VOS to segment multiple objects in one feed-forward path via instance-agnostic and instance-specific modules. Information from the two modules is fused via an attention-guided decoder to simultaneously segment all object instances in one path. The whole framework is end-to-end trainable with instance IoU loss. Experimental results on Youtube- VOS and DAVIS-2017 dataset demonstrate that AGSS-VOS achieves competitive results in terms of both accuracy and efficiency.

****

Global-Local Temporal Representations for Video Person Re-Identification

Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, Shiliang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3958-3967

This paper proposes the Global-Local Temporal Representation (GLTR) to exploit the multi-scale temporal cues in video sequences for video person Re-Identification (ReID). GLTR is constructed by first modeling the short-term temporal cues among adjacent frames, then capturing the long-term relations among inconsecutive frames. Specifically, the short-term temporal cues are modeled by parallel dilated convolutions with different temporal dilation rates to represent the motion and appearance of pedestrian. The long-term relations are captured by a temporal self-attention model to alleviate the occlusions and noises in video sequences. The short and long-term temporal cues are aggregated as the final GLTR by a simple single-stream CNN. GLTR shows substantial superiority to existing features le arned with body part cues or metric learning on four widely-used video ReID datasets. For instance, it achieves Rank-1 Accuracy of 87.02% on MARS dataset without re-ranking, better than current state-of-the art.

****

AdvIT: Adversarial Frames Identifier Based on Temporal Consistency in Videos

Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Benjamin Edwards, Jinfeng Yi, Dawn Song, Mingyan Liu, Ian Molloy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3968-3977

Deep neural networks (DNNs) have been widely applied in various applications, including autonomous driving and surveillance systems. However, DNNs are found to be vulnerable to adversarial examples, which are carefully crafted inputs aiming to mislead a learner to make incorrect predictions. While several defense and detection approaches are proposed for static image classification, many security-critical tasks use videos as their input and require efficient processing. In this paper, we propose an efficient and effective method advIT to detect adversarial frames within videos against different types of attacks based on temporal consistency property of videos. In particular, we apply optical flow estimation to the target and previous frames to generate pseudo frames and evaluate the consistency of the learner output between these pseudo frames and target. High inconsistency indicates that the target frame is adversarial. We conduct extensive experiments on various learning tasks including video semantic segmentation, human pose estimation, object detection, and action recognition, and demonstrate that we can achieve above 95% adversarial frame detection rate. To consider adaptive attackers, we show that even if an adversary has access to the detector and performs a strong adaptive attack based on the state of the art expectation of transformation method, the detection rate stays almost the same. We also tested the transferability among different optical flow estimators and show that it is hard for attackers to attack one and transfer the perturbation to others. In addition, as efficiency is important in video analysis, we show that advIT can achieve real-time detection in about 0.03--0.4 seconds.

****

RANet: Ranking Attention Network for Fast Video Object Segmentation

Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3978-3987

Despite online learning (OL) techniques have boosted the performance of semi-supervised video object segmentation (VOS) methods, the huge time costs of OL greatly restricts their practicality. Matching based and propagation based methods run at a faster speed by avoiding OL techniques. However, they are limited by sub-optimal accuracy, due to mismatching and drifting problems. In this paper, we develop a real-time yet very accurate Ranking Attention Network (RANet) for VOS. Specifically, to integrate the insights of matching based and propagation based methods, we employ an encoder-decoder framework to learn pixel-level similarity and segmentation in an end-to-end manner. To better utilize the similarity maps, we propose a novel ranking attention module, which automatically ranks and selects these maps for fine-grained VOS performance. Experiments on DAVIS16 and DAVIS17 datasets show that our RANet achieves the best speed-accuracy trade-off, e.g., with 33 milliseconds per frame and J&F=85.5% on DAVIS16. With OL, our RANet reaches J&F=87.1% on DAVIS16, exceeding state-of-the-art VOS methods. The code can be found at https://github.com/Storife/RANet.

****

Spatial-Temporal Relation Networks for Multi-Object Tracking

Jiarui Xu, Yue Cao, Zheng Zhang, Han Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3988-3998

Recent progress in multiple object tracking (MOT) has shown that a robust similarity score is a key to the success of trackers. A good similarity score is expected to reflect multiple cues, e.g. appearance, location, and topology, over a long period of time. However, these cues are heterogeneous, making them hard to be combined in a unified network. As a result, existing methods usually encode them in separate networks or require a complex training approach. In this paper, we present a unified framework for similarity measurement based on spatial-tempora l relation network which could simultaneously encode various cues and perform reasoning across both spatial and temporal domains. We also study the feature representation of a tracklet-object pair in depth, showing a proper design of the pair features can well empower the trackers. The resulting approach is named spatial-temporal relation networks (STRN). It runs in a feed-forward way and can be trained in an end-to-end manner. The state-of-the-art accuracy was achieved on all of the MOT15~17 benchmarks using public detection and online settings.

****

Bridging the Gap Between Detection and Tracking: A Unified Approach

Lianghua Huang, Xin Zhao, Kaiqi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3999-4009

Object detection models have been a source of inspiration for many tracking-by-detection algorithms over the past decade. Recent deep trackers borrow designs or modules from the latest object detection methods, such as bounding box regression, RPN and ROI pooling, and can deliver impressive performance. In this paper, instead of redesigning a new tracking-by-detection algorithm, we aim to explore a general framework for building trackers directly upon almost any advanced object detector. To achieve this, three key gaps must be bridged: (1) Object detectors are class-specific, while trackers are class-agnostic. (2) Object detectors do not differentiate intra-class instances, while this is a critical capability of a tracker. (3) Temporal cues are important for stable long-term tracking while they are not considered in still-image detectors. To address the above issues, we first present a simple target-guidance module for guiding the detector to locate target-relevant objects. Then a meta-learner is adopted for the detector to fast learn and adapt a target-distractor classifier online. We further introduce an anchored updating strategy to alleviate the problem of overfitting. The framework is instantiated on SSD and FasterRCNN, the typical one- and two-stage detectors, respectively. Experiments on OTB, UAV123 and NfS have verified our framework and show that our trackers can benefit from deeper backbone networks, as opposed to many recent trackers.

****

Learning the Model Update for Siamese Trackers

Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4010-4019

Siamese approaches address the visual tracking problem by extracting an appearance template from the current frame, which is used to localize the target in the next frame. In general, this template is linearly combined with the accumulated template from the previous frame, resulting in an exponential decay of information over time. While such an approach to updating has led to improved results, its simplicity limits the potential gain likely to be obtained by learning to update. Therefore, we propose to replace the handcrafted update function with a method which learns to update. We use a convolutional neural network, called UpdateNet, which given the initial template, the accumulated template and the template of the current frame aims to estimate the optimal template for the next frame. The UpdateNet is compact and can easily be integrated into existing Siamese trackers. We demonstrate the generality of the proposed approach by applying it to two Siamese trackers, SiamFC and DaSiamRPN. Extensive experiments on VOT2016, VOT2018, LaSOT, and TrackingNet datasets demonstrate that our UpdateNet effectively predicts the new target template, outperforming the standard linear update. On the large-scale TrackingNet dataset, our UpdateNet improves the results of DaSiamRPN with an absolute gain of 3.9% in terms of success score.

****

Fast-deepKCF Without Boundary Effect

Linyu Zheng, Ming Tang, Yingying Chen, Jinqiao Wang, Hanqing Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4020-4029

In recent years, correlation filter based trackers (CF trackers) have received much attention because of their top performance. Most CF trackers, however, suffer from low frame-per-second (fps) in pursuit of higher localization accuracy by relaxing the boundary effect or exploiting the high-dimensional deep features. In order to achieve real-time tracking speed while maintaining high localization accuracy, in this paper, we propose a novel CF tracker, fdKCF*, which casts aside the popular acceleration tool, i.e., fast Fourier transform, employed by all existing CF trackers, and exploits the inherent high-overlap among real (i.e., noncyclic) and dense samples to efficiently construct the kernel matrix. Our fdKCF* enjoys the following three advantages. (i) It is efficiently trained in kernel space and spatial domain without the boundary effect. (ii) Its fps is almost independent of the number of feature channels. Therefore, it is almost real-time, i.e., 24 fps on OTB-2015, even though the high-dimensional deep features are employed. (iii) Its localization accuracy is state-of-the-art. Extensive experiments on four public benchmarks, OTB-2013, OTB-2015, VOT2016, and VOT2017, show that the proposed fdKCF* achieves the state-of-the-art localization performance with remarkably faster speed than C-COT and ECO.

****

Program-Guided Image Manipulators

Jiayuan Mao, Xiuming Zhang, Yikai Li, William T. Freeman, Joshua B. Tenenbaum, Jiajun Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4030-4039

Humans are capable of building holistic representations for images at various levels, from local objects, to pairwise relations, to global structures. The interpretation of structures involves reasoning over repetition and symmetry of the objects in the image. In this paper, we present the Program-Guided Image Manipulator (PG-IM), inducing neuro-symbolic program-like representations to represent and manipulate images. Given an image, PG-IM detects repeated patterns, induces symbolic programs, and manipulates the image using a neural network that is guided by the program. PG-IM learns from a single image, exploiting its internal statistics. Despite trained only on image inpainting, PG-IM is directly capable of extrapolation and regularity editing in a unified framework. Extensive experiments show that PG-IM achieves superior performance on all the tasks.

****

Calibration of Axial Fisheye Cameras Through Generic Virtual Central Models

Pierre-Andre Brousseau, Sebastien Roy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4040-4048

Fisheye cameras are notoriously hard to calibrate using traditional plane-based methods. This paper proposes a new calibration method for large field of view cameras. Similarly to planar calibration, it relies on multiple images of a planar calibration grid with dense correspondences, typically obtained using structured light. By relying on the grids themselves instead of the distorted image plane, we can build a rectilinear Generic Virtual Central (GVC) camera. Instead of relying on a single GVC camera, our method proposes a selection of multiple GVC cameras which can cover any field of view and be trivially aligned to provide a very accurate generic central model. We demonstrate that this approach can directly model axial cameras, assuming the distortion center is located on the camera axis. Experimental validation is provided on both synthetic and real fisheye cameras featuring up to a 280deg field of view. To our knowledge, this is one of the only practical methods to calibrate axial cameras.

****

Micro-Baseline Structured Light

Vishwanath Saragadam, Jian Wang, Mohit Gupta, Shree Nayar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4049-4058

We propose Micro-baseline Structured Light (MSL), a novel 3D imaging approach designed for small form-factor devices such as cell-phones and miniature robots. MSL operates with small projector-camera baseline and low-cost projection hardware, and can recover scene depths with computationally lightweight algorithms. The main observation is that a small baseline leads to small disparities, enabling a first-order approximation of the non-linear SL image formation model. This leads to the key theoretical result of the paper: the MSL equation, a linearized version of SL image formation. MSL equation is under-constrained due to two unknow ns (depth and albedo) at each pixel, but can be efficiently solved using a local least squares approach. We analyze the performance of MSL in terms of various system parameters such as projected pattern and baseline, and provide guidelines for optimizing performance. Armed with these insights, we build a prototype to experimentally examine the theory and its practicality.

****

l-Net: Reconstruct Hyperspectral Images From a Snapshot Measurement

Xin Miao, Xin Yuan, Yunchen Pu, Vassilis Athitsos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4059-4069

We propose the l-net, which reconstructs hyperspectral images (e.g., with 24 spectral channels) from a single shot measurement. This task is usually termed snapshot compressive-spectral imaging (SCI), which enjoys low cost, low bandwidth and high-speed sensing rate via capturing the three-dimensional (3D) signal i.e., (x, y, l), using a 2D snapshot. Though proposed more than a decade ago, the poor quality and low-speed of reconstruction algorithms preclude wide applications of SCI. To address this challenge, in this paper, we develop a dual-stage generative model to reconstruct the desired 3D signal in SCI, dubbed l-net. Results on both simulation and real datasets demonstrate the significant advantages of l-net, which leads to >4dB improvement in PSNR for real-mask-in-the-loop simulation data compared to the current state-of-the-art. Furthermore, l-net can finish the reconstruction task within sub-seconds instead of hours taken by the most recently proposed DeSCI algorithm, thus speeding up the reconstruction >1000 times.

****

Deep Depth From Aberration Map

Masako Kashiwagi, Nao Mishima, Tatsuo Kozakaya, Shinsaku Hiura; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4070-4079

Passive and convenient depth estimation from single-shot image is still an open problem. Existing depth from defocus methods require multiple input images or special hardware customization. Recent deep monocular depth estimation is also limited to an image with sufficient contextual information. In this work, we propose a novel method which realizes a single-shot deep depth measurement based on physical depth cue using only an off-the-shelf camera and lens. When a defocused image is taken by a camera, it contains various types of aberrations corresponding to distances from the image sensor and positions in the image plane. We call these minute and complexly compound aberrations as Aberration Map (A-Map) and we found that A-Map can be utilized as reliable physical depth cue. Additionally, our deep network named A-Map Analysis Network (AMA-Net) is also proposed, which can effectively learn and estimate depth via A-Map. To evaluate validity and robustness of our approach, we have conducted extensive experiments using both real outdoor scenes and simulated images. The qualitative result shows the accuracy and availability of the method in comparison with a state-of-the-art deep context-based method.

****

A Dataset of Multi-Illumination Images in the Wild

Lukas Murmann, Michael Gharbi, Miika Aittala, Fredo Durand; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4080-4089

Collections of images under a single, uncontrolled illumination have enabled the rapid advancement of core computer vision tasks like classification, detection, and segmentation. But even with modern learning techniques, many inverse problems involving lighting and material understanding remain too severely ill-posed to be solved with single-illumination datasets. The data simply does not contain the necessary supervisory signals. Multi-illumination datasets are notoriously hard to capture, so the data is typically collected at small scale, in controlled environments, either using multiple light sources, or robotic gantries. This leads to image collections that are not representative of the variety and complexity of real world scenes. We introduce a new multi-illumination dataset of more than 1000 real scenes, each captured in high dynamic range and high resolution, under 25 lighting conditions. We demonstrate the richness of this dataset by trai ning state-of-the-art models for three challenging applications: single-image illumination estimation, image relighting, and mixed-illuminant white balance.

****

Monocular Neural Image Based Rendering With Continuous View Control

Xu Chen, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4090-4100

We propose a method to produce a continuous stream of novel views under fine-grained (e.g., 1 degree step-size) camera control at interactive rates. A novel learning pipeline determines the output pixels directly from the source color. Injecting geometric transformations, including perspective projection, 3D rotation and translation into the network forces implicit reasoning about the underlying geometry. The latent 3D geometry representation is compact and meaningful under 3D transformation, being able to produce geometrically accurate views for both single objects and natural scenes. Our experiments show that both proposed components, the transforming encoder-decoder and depth-guided appearance mapping, lead to significantly improved generalization beyond the training views and in consequence to more accurate view synthesis under continuous 6-DoF camera control. Finally, we show that our method outperforms state-of-the-art baseline methods on public datasets.

****

Multi-View Image Fusion

Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, Janne Kontkanen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4101-4110

We present an end-to-end learned system for fusing multiple misaligned photographs of the same scene into a chosen target view. We demonstrate three use cases: 1) color transfer for inferring color for a monochrome view, 2) HDR fusion for merging misaligned bracketed exposures, and 3) detail transfer for reprojecting a high definition image to the point of view of an affordable VR180-camera. While the system can be trained end-to-end, it consists of three distinct steps: feature extraction, image warping and fusion. We present a novel cascaded feature extraction method that enables us to synergetically learn optical flow at different resolution levels. We show that this significantly improves the network's ability to learn large disparities. Finally, we demonstrate that our alignment architecture outperforms a state-of-the art optical flow network on the image warping task when both systems are trained in an identical manner.

****

Enhancing Low Light Videos by Exploring High Sensitivity Camera Noise

Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, Tao Yue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4111-4119

Enhancing low light videos, which consists of denoising and brightness adjustment, is an intriguing but knotty problem. Under low light condition, due to high sensitivity camera setting, commonly negligible noises become obvious and severely deteriorate the captured videos. To recover high quality videos, a mass of image/video denoising/enhancing algorithms are proposed, most of which follow a set of simple assumptions about the statistic characters of camera noise, e.g., independent and identically distributed(i.i.d.), white, additive, Gaussian, Poisson or mixture noises. However, the practical noise under high sensitivity setting in real captured videos is complex and inaccurate to model with these assumptions. In this paper, we explore the physical origins of the practical high sensitivity noise in digital cameras, model them mathematically, and propose to enhance the low light videos based on the noise model by using an LSTM-based neural network. Specifically, we generate the training data with the proposed noise model and train the network with the dark noisy video as input and clear-bright video as output. Extensive comparisons on both synthetic and real captured low light videos with the state-of-the-art methods are conducted to demonstrate the effectiveness of the proposed method.

****

Deep Restoration of Vintage Photographs From Scanned Halftone Prints

Qifan Gao, Xiao Shu, Xiaolin Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4120-4129

A great number of invaluable historical photographs unfortunately only exist in the form of halftone prints in old publications such as newspapers or books. Their original continuous-tone films have long been lost or irreparably damaged. There have been attempts to digitally restore these vintage halftone prints to the original film quality or higher. However, even using powerful deep convolutional neural networks, it is still difficult to obtain satisfactory results. The main challenge is that the degradation process is complex and compounded while little to no real data is available for properly training a data-driven method. In this research, we adopt a novel strategy of two-stage deep learning, in which the restoration task is divided into two stages: the removal of printing artifacts and the inverse of halftoning. The advantage of our technique is that only the simple first stage requires unsupervised training in order to make the combined network generalize on real halftone prints, while the more complex second stage of inverse halftoning can be easily trained with synthetic data. Extensive experimental results demonstrate the efficacy of the proposed technique for real halftone prints; the new technique significantly outperforms the existing ones in visual quality.

****

Context-Aware Image Matting for Simultaneous Foreground and Alpha Estimation

Qiqi Hou, Feng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4130-4139

Natural image matting is an important problem in computer vision and graphics. It is an ill-posed problem when only an input image is available without any external information. While the recent deep learning approaches have shown promising results, they only estimate the alpha matte. This paper presents a context-aware natural image matting method for simultaneous foreground and alpha matte estimation. Our method employs two encoder networks to extract essential information for matting. Particularly, we use a matting encoder to learn local features and a context encoder to obtain more global context information. We concatenate the outputs from these two encoders and feed them into decoder networks to simultaneously estimate the foreground and alpha matte. To train this whole deep neural network, we employ both the standard Laplacian loss and the feature loss: the former helps to achieve high numerical performance while the latter leads to more perceptually plausible results. We also report several data augmentation strategies that greatly improve the network's generalization performance. Our qualitative and quantitative experiments show that our method enables high-quality matting for a single natural image.

****

CFSNet: Toward a Controllable Feature Space for Image Restoration

Wei Wang, Ruiming Guo, Yapeng Tian, Wenming Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4140-4149

Deep learning methods have witnessed the great progress in image restoration with specific metrics (e.g., PSNR, SSIM). However, the perceptual quality of the restored image is relatively subjective, and it is necessary for users to control the reconstruction result according to personal preferences or image characteristics, which cannot be done using existing deterministic networks. This motivates us to exquisitely design a unified interactive framework for general image restoration tasks. Under this framework, users can control continuous transition of different objectives, e.g., the perception-distortion trade-off of image super-resolution, the trade-off between noise reduction and detail preservation. We achieve this goal by controlling the latent features of the designed network. To be specific, our proposed framework, named Controllable Feature Space Network (CFSNet), is entangled by two branches based on different objectives. Our framework can adaptively learn the coupling coefficients of different layers and channels, which provides finer control of the restored image quality. Experiments on several typical image restoration tasks fully validate the effective benefits of the proposed method. Code is available at https://github.com/qibao77/CFSNet.

****

Deep Blind Hyperspectral Image Fusion

Wu Wang, Weihong Zeng, Yue Huang, Xinghao Ding, John Paisley; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4150-4159

Hyperspectral image fusion (HIF) reconstructs high spatial resolution hyperspectral images from low spatial resolution hyperspectral images and high spatial resolution multispectral images. Previous works usually assume that the linear mapping between the point spread functions of the hyperspectral camera and the spectral response functions of the conventional camera is known. This is unrealistic in many scenarios. We propose a method for blind HIF problem based on deep learning, where the estimation of the observation model and fusion process are optimized iteratively and alternatingly during the super-resolution reconstruction. In addition, the proposed framework enforces simultaneous spatial and spectral accuracy. Using three public datasets, the experimental results demonstrate that the proposed algorithm outperforms existing blind and non-blind methods.

****

Fully Convolutional Pixel Adaptive Image Denoiser

Sungmin Cha, Taesup Moon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4160-4169

We propose a new image denoising algorithm, dubbed as Fully Convolutional Adaptive Image DEnoiser (FC-AIDE), that can learn from an offline supervised training set with a fully convolutional neural network as well as adaptively fine-tune the supervised model for each given noisy image. We significantly extend the framework of the recently proposed Neural AIDE, which formulates the denoiser to be context-based pixelwise mappings and utilizes the unbiased estimator of MSE for such denoisers. The two main contributions we make are; 1) implementing a novel fully convolutional architecture that boosts the base supervised model, and 2) introducing regularization methods for the adaptive fine-tuning such that a stronger and more robust adaptivity can be attained. As a result, FC-AIDE is shown to possess many desirable features; it outperforms the recent CNN-based state-of-the-art denoisers on all of the benchmark datasets we tested, and gets particularly strong for various challenging scenarios, e.g., with mismatched image/noise characteristics or with scarce supervised training data. The source code our algorithm is available at https://github.com/csm9493/FC-AIDE-Keras https://github.com/csm9493/FC-AIDE-Keras .

****

Coherent Semantic Attention for Image Inpainting

Hongyu Liu, Bin Jiang, Yi Xiao, Chao Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4170-4179

The latest deep learning-based approaches have shown promising results for the challenging task of inpainting missing regions of an image. However, the existing methods often generate contents with blurry textures and distorted structures due to the discontinuity of the local pixels. From a semantic-level perspective, the local pixel discontinuity is mainly because these methods ignore the semantic relevance and feature continuity of hole regions. To handle this problem, we investigate the human behavior in repairing pictures and propose a fined deep generative model-based approach with a novel coherent semantic attention (CSA) layer, which can not only preserve contextual structure but also make more effective predictions of missing parts by modeling the semantic relevance between the holes features. The task is divided into rough, refinement as two steps and we model each step with a neural network under the U-Net architecture, where the CSA layer is embedded into the encoder of refinement step. Meanwhile, we further propose consistency loss and feature patch discriminator to stabilize the network training process and improve the details. The experiments on CelebA, Places2, and Paris StreetView datasets have validated the effectiveness of our proposed methods in image inpainting tasks and can obtain images with a higher quality as compared with the existing state-of-the-art approaches. The codes and pre-trained models will be available at https://github.com/KumapowerLIU/CSA-inpainting.

****

Embedded Block Residual Network: A Recursive Restoration Model for Single-Image

Super-Resolution

Yajun Qiu, Ruxin Wang, Dapeng Tao, Jun Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4180-4189

Single-image super-resolution restores the lost structures and textures from low-resolved images, which has achieved extensive attention from the research community. The top performers in this field include deep or wide convolutional neural networks, or recurrent neural networks. However, the methods enforce a single model to process all kinds of textures and structures. A typical operation is that a certain layer restores the textures based on the ones recovered by the preceding layers, ignoring the characteristics of image textures. In this paper, we believe that the lower-frequency and higher-frequency information in images have different levels of complexity and should be restored by models of different representational capacity. Inspired by this, we propose a novel embedded block residual network (EBRN) which is an incremental recovering progress for texture super-resolution. Specifically, different modules in the model restores information of different frequencies. For lower-frequency information, we use shallower modules of the network to recover; for higher-frequency information, we use deeper modules to restore. Extensive experiments indicate that the proposed EBRN model achieves superior performance and visual improvements against the state-of-the-arts.

****

Fast Image Restoration With Multi-Bin Trainable Linear Units

Shuhang Gu, Wen Li, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4190-4199

Tremendous advances in image restoration tasks such as denoising and super-resolution have been achieved using neural networks. Such approaches generally employ very deep architectures, large number of parameters, large receptive fields and high nonlinear modeling capacity. In order to obtain efficient and fast image restoration networks one should improve upon the above mentioned requirements. In this paper we propose a novel activation function, the multi-bin trainable linear unit (MTLU), for increasing the nonlinear modeling capacity together with lighter and shallower networks. We validate the proposed fast image restoration networks for image denoising (FDnet) and super-resolution (FSRnet) on standard benchmarks. We achieve large improvements in both memory and runtime over current state-of-the-art for comparable or better PSNR accuracies.

****

Counting With Focus for Free

Zenglin Shi, Pascal Mettes, Cees G. M. Snoek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4200-4209

This paper aims to count arbitrary objects in images. The leading counting approaches start from point annotations per object from which they construct density maps. Then, their training objective transforms input images to density maps through deep convolutional networks. We posit that the point annotations serve more supervision purposes than just constructing density maps. We introduce ways to repurpose the points for free. First, we propose supervised focus from segmentation, where points are converted into binary maps. The binary maps are combined with a network branch and accompanying loss function to focus on areas of interest. Second, we propose supervised focus from global density, where the ratio of point annotations to image pixels is used in another branch to regularize the overall density estimation. To assist both the density estimation and the focus from segmentation, we also introduce an improved kernel size estimator for the point annotations. Experiments on six datasets show that all our contributions reduce the counting error, regardless of the base network, resulting in state-of-the-art accuracy using only a single network. Finally, we are the first to count on WIDER FACE, allowing us to show the benefits of our approach in handling varying object scales and crowding levels. Code is available at https://github.com/shizenglin/Counting-with-Focus-for-Free

****

SynDeMo: Synergistic Deep Feature Alignment for Joint Learning of Depth and Ego-Motion

Behzad Bozorgtabar, Mohammad Saeed Rad, Dwarikanath Mahapatra, Jean-Philippe Thiran; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4210-4219

Despite well-established baselines, learning of scene depth and ego-motion from monocular video remains an ongoing challenge, specifically when handling scaling ambiguity issues and depth inconsistencies in image sequences. Much prior work uses either a supervised mode of learning or stereo images. The former is limited by the amount of labeled data, as it requires expensive sensors, while the latter is not always readily available as monocular sequences. In this work, we demonstrate the benefit of using geometric information from synthetic images, coupled with scene depth information, to recover the scale in depth and ego-motion estimation from monocular videos. We developed our framework using synthetic image-depth pairs and unlabeled real monocular images. We had three training objectives: first, to use deep feature alignment to reduce the domain gap between synthetic and monocular images to yield more accurate depth estimation when presented with only real monocular images at test time. Second, we learn scene specific representation by exploiting self-supervision coming from multi-view synthetic images without the need for depth labels. Third, our method uses single-view depth and pose networks, which are capable of jointly training and supervising one another mutually, yielding consistent depth and ego-motion estimates. Extensive experiments demonstrate that our depth and ego-motion models surpass the state-of-the-art, unsupervised methods and compare favorably to early supervised deep models for geometric understanding. We validate the effectiveness of our training objectives against standard benchmarks thorough an ablation study.

****

Diverse Image Synthesis From Semantic Layouts via Conditional IMLE

Ke Li, Tianhao Zhang, Jitendra Malik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4220-4229

Most existing methods for conditional image synthesis are only able to generate a single plausible image for any given input, or at best a fixed number of plausible images. In this paper, we focus on the problem of generating images from semantic segmentation maps and present a simple new method that can generate an arbitrary number of images with diverse appearance for the same semantic layout. Unlike most existing approaches which adopt the GAN framework, our method is based on the recently introduced Implicit Maximum Likelihood Estimation (IMLE) framework. Compared to the leading approach, our method is able to generate more diverse images while producing fewer artifacts despite using the same architecture. The learned latent space also has sensible structure despite the lack of supervision that encourages such behaviour.

****

Towards Bridging Semantic Gap to Improve Semantic Segmentation

Yanwei Pang, Yazhao Li, Jianbing Shen, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4230-4239

Aggregating multi-level features is essential for capturing multi-scale context information for precise scene semantic segmentation. However, the improvement by directly fusing shallow features and deep features becomes limited as the semantic gap between them increases. To solve this problem, we explore two strategies for robust feature fusion. One is enhancing shallow features using a semantic enhancement module (SeEM) to alleviate the semantic gap between shallow features and deep features. The other strategy is feature attention, which involves discovering complementary information (i.e., boundary information) from low-level features to enhance high-level features for precise segmentation. By embedding these two strategies, we construct a parallel feature pyramid towards improving multi-level feature fusion. A Semantic Enhanced Network called SeENet is constructed with the parallel pyramid to implement precise segmentation. Experiments on three benchmark datasets demonstrate the effectiveness of our method for robust multi-level feature aggregation. As a result, our SeENet has achieved better performance than other state-of-the-art methods for semantic segmentation.

****

Generating Diverse and Descriptive Image Captions Using Visual Paraphrases

Lixin Liu, Jiajun Tang, Xiaojun Wan, Zongming Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4240-4249

Recently there has been significant progress in image captioning with the help of deep learning. However, captions generated by current state-of-the-art models are still far from satisfactory, despite high scores in terms of conventional metrics such as BLEU and CIDEr. Human-written captions are diverse, informative and precise, but machine-generated captions seem to be simple, vague and dull. In this paper, aimed at improving diversity and descriptiveness characteristics of generated image captions, we propose a model utilizing visual paraphrases (different sentences describing the same image) in captioning datasets. We explore different strategies to select useful visual paraphrase pairs for training by designing a variety of scoring functions. Our model consists of two decoding stages, where a preliminary caption is generated in the first stage and then paraphrased into a more diverse and descriptive caption in the second stage. Extensive experiments are conducted on the benchmark MS COCO dataset, with automatic evaluation and human evaluation results verifying the effectiveness of our model.

****

Learning to Collocate Neural Modules for Image Captioning

Xu Yang, Hanwang Zhang, Jianfei Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4250-4260

We do not speak word by word from scratch; our brain quickly structures a pattern like sth do sth at someplace and then fill in the detailed description. To render existing encoder-decoder image captioners such human-like reasoning, we propose a novel framework: learning to Collocate Neural Modules (CNM), to generate the "inner pattern" connecting visual encoder and language decoder. Unlike the widely-used neural module networks in visual Q&A, where the language (i.e., question) is fully observable, CNM for captioning is more challenging as the language is being generated and thus is partially observable. To this end, we make the following technical contributions for CNM training: 1) compact module design --- one for function words and three for visual content words (e.g., noun, adjective, and verb), 2) soft module fusion and multi-step module execution, robustifying the visual reasoning in partial observation, 3) a linguistic loss for module controller being faithful to part-of-speech collocations (e.g., adjective is before noun). Extensive experiments on the challenging MS-COCO image captioning benchmark validate the effectiveness of our CNM image captioner. In particular, CNM achieves a new state-of-the-art 127.9 CIDEr-D on Karpathy split and a single-model 126.0 c40 on the official server. CNM is also robust to few training samples, e.g., by training only one sentence per image, CNM can halve the performance loss compared to a strong baseline.

****

Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, Alexander Schwing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4261-4270

Diverse and accurate vision+language modeling is an important goal to retain creative freedom and maintain user engagement. However, adequately capturing the intricacies of diversity in language models is challenging. Recent works commonly resort to latent variable models augmented with more or less supervision from object detectors or part-of-speech tags. In common to all those methods is the fact that the latent variable either only initializes the sentence generation process or is identical across the steps of generation. Both methods offer no fine-grained control. To address this concern, we propose Seq-CVAE which learns a latent space for every word. We encourage this temporal latent space to capture the 'intention' about how to complete the sentence by mimicking a representation which summarizes the future. We illustrate the efficacy of the proposed approach on the challenging MSCOCO dataset, significantly improving diversity metrics compared to baselines while performing on par w.r.t. sentence quality.

****

Why Does a Visual Question Have Different Answers?

Nilavra Bhattacharya, Qing Li, Danna Gurari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4271-4280

Visual question answering is the task of returning the answer to a question about an image. A challenge is that different people often provide different answers to the same visual question. To our knowledge, this is the first work that aims to understand why. We propose a taxonomy of nine plausible reasons, and create two labelled datasets consisting of 45,000 visual questions indicating which reasons led to answer differences. We then propose a novel problem of predicting directly from a visual question which reasons will cause answer differences as well as a novel algorithm for this purpose. Experiments demonstrate the advantage of our approach over several related baselines on two diverse datasets. We publicly share the datasets and code at https://vizwiz.org.

****

G3raphGround: Graph-Based Language Grounding

Mohit Bajaj, Lanjun Wang, Leonid Sigal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4281-4290

In this paper we present an end-to-end framework for grounding of phrases in images. In contrast to previous works, our model, which we call GraphGround, uses graphs to formulate more complex, non-sequential dependencies among proposal image regions and phrases. We capture intra-modal dependencies using a separate graph neural network for each modality (visual and lingual), and then use conditional message-passing in another graph neural network to fuse their outputs and capture cross-modal relationships. This final representation results in grounding decisions. The framework supports many-to-many matching and is able to ground single phrase to multiple image regions and vice versa. We validate our design choices through a series of ablation studies and illustrate state-of-the-art performance on Flickr30k and ReferIt Game benchmark datasets.

****

Scene Text Visual Question Answering

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, Dimosthenis Karatzas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4291-4301

Current visual question answering datasets do not consider the rich semantic information conveyed by text within an image. In this work, we present a new dataset, ST-VQA, that aims to highlight the importance of exploiting high-level semantic information present in images as textual cues in the Visual Question Answering process. We use this dataset to define a series of tasks of increasing difficulty for which reading the scene text in the context provided by the visual information is necessary to reason and generate an appropriate answer. We propose a new evaluation metric for these tasks to account both for reasoning errors as well as shortcomings of the text recognition module. In addition we put forward a series of baseline methods, which provide further insight to the newly released dataset, and set the scene for further research.

****

Unsupervised Collaborative Learning of Keyframe Detection and Visual Odometry Towards Monocular Deep SLAM

Lu Sheng, Dan Xu, Wanli Ouyang, Xiaogang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4302-4311

In this paper we tackle the joint learning problem of keyframe detection and visual odometry towards monocular visual SLAM systems. As an important task in visual SLAM, keyframe selection helps efficient camera relocalization and effective augmentation of visual odometry. To benefit from it, we first present a deep network design for the keyframe selection, which is able to reliably detect keyframes and localize new frames, then an end-to-end unsupervised deep framework further proposed for simultaneously learning the keyframe selection and the visual odometry tasks. As far as we know, it is the first work to jointly optimize these two complementary tasks in a single deep framework. To make the two tasks facilitate each other in the learning, a collaborative optimization loss based on both geometric and visual metrics is proposed. Extensive experiments on publicly available datasets (i.e. KITTI raw dataset and its odometry split) clearly demonstr ate the effectiveness of the proposed approach, and new state-of-the-art results are established on the unsupervised depth and pose estimation from monocular videos.

****

MVSCRF: Learning Multi-View Stereo With Conditional Random Fields

Youze Xue, Jiansheng Chen, Weitao Wan, Yiqing Huang, Cheng Yu, Tianpeng Li, Jiayu Bao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4312-4321

We present a deep-learning architecture for multi-view stereo with conditional random fields (MVSCRF). Given an arbitrary number of input images, we first use a U-shape neural network to extract deep features incorporating both global and local information, and then build a 3D cost volume for the reference camera. Unlike previous learning based methods, we explicitly constraint the smoothness of depth maps by using conditional random fields (CRFs) after the stage of cost volume regularization. The CRFs module is implemented as recurrent neural networks so that the whole pipeline can be trained end-to-end. Our results show that the proposed pipeline outperforms previous state-of-the-arts on large-scale DTU dataset. We also achieve comparable results with state-of-the-art learning based methods on outdoor Tanks and Temples dataset without fine-tuning, which demonstrates our method's generalization ability.

****

Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses

Eric Brachmann, Carsten Rother; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4322-4331

We present Neural-Guided RANSAC (NG-RANSAC), an extension to the classic RANSAC algorithm from robust optimization. NG-RANSAC uses prior information to improve model hypothesis search, increasing the chance of finding outlier-free minimal sets. Previous works use heuristic side-information like hand-crafted descriptor distance to guide hypothesis search. In contrast, we learn hypothesis search in a principled fashion that lets us optimize an arbitrary task loss during training, leading to large improvements on classic computer vision tasks. We present two further extensions to NG-RANSAC. Firstly, using the inlier count itself as training signal allows us to train neural guidance in a self-supervised fashion. Secondly, we combine neural guidance with differentiable RANSAC to build neural networks which focus on certain parts of the input data and make the output predictions as good as possible. We evaluate NG-RANSAC on a wide array of computer vision tasks, namely estimation of epipolar geometry, horizon line estimation and camera re-localization. We achieve superior or competitive results compared to state-of-the-art robust estimators, including very recent, learned ones.

****

Efficient Learning on Point Clouds With Basis Point Sets

Sergey Prokudin, Christoph Lassner, Javier Romero; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4332-4341

With an increased availability of 3D scanning technology, point clouds are moving into the focus of computer vision as a rich representation of everyday scenes. However, they are hard to handle for machine learning algorithms due to the unordered structure. One common approach is to apply voxelization, which dramatically increases the amount of data stored and at the same time loses details through discretization. Recently, deep learning models with hand-tailored architectures were proposed to handle point clouds directly and achieve input permutation invariance. However, these architectures use an increased number of parameters and are computationally inefficient. In this work we propose basis point sets as a highly efficient and fully general way to process point clouds with machine learning algorithms. Basis point sets are a residual representation that can be computed efficiently and can be used with standard neural network architectures. Using the proposed representation as the input to a relatively simple network allows us to match the performance of PointNet on a shape classification task while using three order of magnitudes less floating point operations. In a second experiment, we show how proposed representation can be used for obtaining high resolution meshes from noisy 3D scans. Here, our network achieves performance comparab le to the state-of-the-art computationally intense multi-step frameworks, in one network pass that can be done in less than 1ms.

****

Cross View Fusion for 3D Human Pose Estimation

Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, Wenjun Zeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4342-4351

We present an approach to recover absolute 3D human poses from multi-view images by incorporating multi-view geometric priors in our model. It consists of two separate steps: (1) estimating the 2D poses in multi-view images and (2) recovering the 3D poses from the multi-view 2D poses. First, we introduce a cross-view fusion scheme into CNN to jointly estimate 2D poses for multiple views. Consequently, the 2D pose estimation for each view already benefits from other views. Second, we present a recursive Pictorial Structure Model to recover the 3D pose from the multi-view 2D poses. It gradually improves the accuracy of 3D pose with affordable computational cost. We test our method on two public datasets H36M and Total Capture. The Mean Per Joint Position Errors on the two datasets are 26mm and 29mm, which outperforms the state-of-the-arts remarkably (26mm vs 52mm, 29mm vs 35mm).

****

Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images

Junbang Liang, Ming C. Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4352-4362

We propose a scalable neural network framework to reconstruct the 3D mesh of a human body from multi-view images, in the subspace of the SMPL model. Use of multi-view images can significantly reduce the projection ambiguity of the problem, increasing the reconstruction accuracy of the 3D human body under clothing. Our experiments show that this method benefits from the synthetic dataset generated from our pipeline since it has good flexibility of variable control and can provide ground-truth for validation. Our method outperforms existing methods on real-world images, especially on shape estimations.

****

Monocular Piecewise Depth Estimation in Dynamic Scenes by Exploiting Superpixel Relations

Yan Di, Henrique Morimitsu, Shan Gao, Xiangyang Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4363-4372

In this paper, we propose a novel and specially designed method for piecewise dense monocular depth estimation in dynamic scenes. We utilize spatial relations between neighboring superpixels to solve the inherent relative scale ambiguity (RSA) problem and smooth the depth map. However, directly estimating spatial relations is an ill-posed problem. Our core idea is to predict spatial relations based on the corresponding motion relations. Given two or more consecutive frames, we first compute semi-dense (CPM) or dense (optical flow) point matches between temporally neighboring images. Then we develop our method in four main stages: superpixel relations analysis, motion selection, reconstruction, and refinement. The final refinement process helps to improve the quality of the reconstruction at pixel level. Our method does not require per-object segmentation, template priors or training sets, which ensures flexibility in various applications. Extensive experiments on both synthetic and real datasets demonstrate that our method robustly handles different dynamic situations and presents competitive results to the state-of-the-art methods while running much faster than them.

****

Is This the Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization

Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, Akihiko Torii; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4373-4383

Visual localization in large and complex indoor scenes, dominated by weakly textured rooms and repeating geometric patterns, is a challenging problem with high practical relevance for applications such as Augmented Reality and robotics. To handle the ambiguities arising in this scenario, a common strategy is, first, to generate multiple estimates for the camera pose from which a given query image was taken. The pose with the largest geometric consistency with the query image, e.g., in the form of an inlier count, is then selected in a second stage. While a significant amount of research has concentrated on the first stage, there has been considerably less work on the second stage. In this paper, we thus focus on pose verification. We show that combining different modalities, namely appearance, geometry, and semantics, considerably boosts pose verification and consequently pose accuracy. We develop multiple hand-crafted as well as a trainable approach to join into the geometric-semantic verification and show significant improvements over state-of-the-art on a very challenging indoor dataset.

****

DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch

Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4384-4393

Our goal is to significantly speed up the runtime of current state-of-the-art stereo algorithms to enable real-time inference. Towards this goal, we developed a differentiable PatchMatch module that allows us to discard most disparities without requiring full cost volume evaluation. We then exploit this representation to learn which range to prune for each pixel. By progressively reducing the search space and effectively propagating such information, we are able to efficiently compute the cost volume for high likelihood hypotheses and achieve savings in both memory and computation.Finally, an image guided refinement module is exploited to further improve the performance. Since all our components are differentiable, the full network can be trained end-to-end. Our experiments show that our method achieves competitive results on KITTI and SceneFlow datasets while running in real-time at 62ms.

****

Convolutional Sequence Generation for Skeleton-Based Action Synthesis

Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4394-4402

In this work, we aim to generate long actions represented as sequences of skeletons. The generated sequences must demonstrate continuous, meaningful human actions, while maintaining coherence among body parts. Instead of generating skeletons sequentially following an autoregressive model, we propose a framework that generates the entire sequence altogether by transforming from a sequence of latent vectors sampled from a Gaussian process (GP). This framework, named Convolutional Sequence Generation Network (CSGN), jointly models structures in temporal and spatial dimensions. It captures the temporal structure at multiple scales through the GP prior and the temporal convolutions; and establishes the spatial connection between the latent vectors and the skeleton graphs via a novel graph refining scheme. It is noteworthy that CSGN allows bidirectional transforms between the latent and the observed spaces, thus enabling semantic manipulation of the action sequences in various forms. We conducted empirical studies on multiple datasets, including a set of high-quality dancing sequences collected by us. The results show that our framework can produce long action sequences that are coherent across time steps and among body parts.

****

Onion-Peel Networks for Deep Video Completion

Seoung Wug Oh, Sungho Lee, Joon-Young Lee, Seon Joo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4403-4412

We propose the onion-peel networks for video completion. Given a set of reference images and a target image with holes, our network fills the hole by referring the contents in the reference images. Our onion-peel network progressively fills the hole from the hole boundary enabling it to exploit richer contextual information for the missing regions every step. Given a sufficient number of recurrences, even a large hole can be inpainted successfully. To attend to the missing information visible in the reference images, we propose an asymmetric attention bl ock that computes similarities between the hole boundary pixels in the target and the non-hole pixels in the references in a non-local manner. With our attention block, our network can have an unlimited spatial-temporal window size and fill the holes with globally coherent contents. In addition, our framework is applicable to the image completion guided by the reference images without any modification, which is difficult to do with the previous methods. We validate that our method produces visually pleasing image and video inpainting results in realistic test cases.

****

Copy-and-Paste Networks for Deep Video Inpainting

Sungho Lee, Seoung Wug Oh, DaeYeun Won, Seon Joo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4413-4421

We present a novel deep learning based algorithm for video inpainting. Video inpainting is a process of completing corrupted or missing regions in videos. Video inpainting has additional challenges compared to image inpainting due to the extra temporal information as well as the need for maintaining the temporal coherency. We propose a novel DNN-based framework called the Copy-and-Paste Networks for video inpainting that takes advantage of additional information in other frames of the video. The network is trained to copy corresponding contents in reference frames and paste them to fill the holes in the target frame. Our network also includes an alignment network that computes homographies between frames for the alignment, enabling the network to take information from more distant frames for robustness. Our method produces visually pleasing and temporally coherent results while running faster than the state-of-the-art optimization-based method. In addition, we extend our framework for enhancing over/under exposed frames in videos. Using this enhancement technique, we were able to significantly improve the lane detection accuracy on road videos.

****

Content and Style Disentanglement for Artistic Style Transfer

Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, Bjorn Ommer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4422-4431

Artists rarely paint in a single style throughout their career. More often they change styles or develop variations of it. In addition, artworks in different styles and even within one style depict real content differently: while Picasso's Blue Period displays a vase in a blueish tone but as a whole, his Cubist works deconstruct the object. To produce artistically convincing stylizations, style transfer models must be able to reflect these changes and variations. Recently many works have aimed to improve the style transfer task, but neglected to address the described observations. We present a novel approach which captures particularities of style and the variations within and separates style and content. This is achieved by introducing two novel losses: a fixpoint triplet style loss to learn subtle variations within one style or between different styles and a disentanglement loss to ensure that the stylization is not conditioned on the real input photo. In addition the paper proposes various evaluation methods to measure the importance of both losses on the validity, quality and variability of final stylizations. We provide qualitative results to demonstrate the performance of our approach.

****

Compositional Video Prediction

Yufei Ye, Maneesh Singh, Abhinav Gupta, Shubham Tulsiani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10353-10362

We present an approach for pixel-level future prediction given an input image of a scene. We observe that a scene is comprised of distinct entities that undergo motion and present an approach that operationalizes this insight. We implicitly predict future states of independent entities while reasoning about their interactions, and compose future video frames using these predicted states. We overcome the inherent multi-modality of the task using a global trajectory-level latent random variable, and show that this allows us to sample diverse and plausible futures. We empirically validate our approach against alternate representations and ways of incorporating multi-modality. We examine two datasets, one comprising of stacked objects that may fall, and the other containing videos of humans performing activities in a gym, and show that our approach allows realistic stochastic video prediction across these diverse settings. See project website (https://judyye.github.io/CVP/) for video predictions.

****